

*AN INVESTIGATION INTO THE RATER
COGNITION OF NOVICE RATERS AND THE
IMPACT OF COGNITIVE ATTRIBUTES WHEN
ASSESSING SPEAKING*



Mag. phil. Kathrin Eberharter, MA

**Thesis submitted to Lancaster University in fulfilment of the requirements for the
degree of Doctor of Philosophy**

July 2021

This dissertation is dedicated to my late grandmother,

Maria Eberharter,

for always seeing the explorer in me.

ABSTRACT

Examinations of language proficiency routinely include the assessment of speaking, which still largely necessitates the use of human raters. However, variability in rating quality is a well-established phenomenon and makes rating a fundamental validity concern (Kane, 1992; 2006). Despite increased efforts to investigate rater cognition to better understand and mitigate rater effects (Bejar, 2012), research in language testing is yet to fully engage with the field of decision research (Baker, 2012; Purpura, 2013). Findings from this literature emphasize how complex decision tasks are shaped by factors such as processing capacities, perception, deliberate and automated thinking, and metacognitive control (Newell and Bröder, 2008).

The purpose of this study was to investigate how novice raters use an analytic rating scale and to explore whether decision-making style, cognitive style, working memory capacity and executive function influence rating quality and rating behaviour. 39 pre-service English teachers rated a set of speaking performances ($N=30$) and completed two psychological questionnaires as well as a battery of cognitive tests. Rating behaviours were captured through JavaScript embedded in the online rating form.

Data analysis first established measures of rating quality and scale use through a series of Many-Facets Rasch Measurement (MFRM) analyses. Next, relationships between individual attributes and measures of rater quality and behaviour were explored in a series of correlational analyses. Finally, the handwritten notes and self-report data from four selected raters were accumulated and explored to further enhance understanding of the rating process.

Findings showed that there were considerable individual differences among the raters regarding rating quality and behaviours. Of all the variables included, decision-making style displayed the strongest associations with rating quality and behaviour, suggesting a

relationship between intuitive and flexible processing and more successful rating. The four case studies highlighted a need to address cognitive load and directing of attention in rater training for speaking assessment.

DECLARATION

I declare that this thesis is the result of my own research except as cited in references.
The thesis has not been submitted in candidature of any other degree.

Signed: 

ACKNOWLEDGEMENTS

Like many others who embarked on the endeavour of obtaining their doctorate degree I realized quickly that it takes a village to raise a PhD. I was fortunate to have received a great deal of support and assistance throughout the writing of this dissertation.

I would like to express my great appreciation to my fantastic supervisor, Professor Luke Harding for his patience, enthusiastic encouragement and guidance. Without his continuous optimism and constructive thinking, this study would simply not have been completed.

This dissertation would also not have been possible without the support of Univ.-Prof. Dr. Barbara Hinger MA and Univ.-Prof. Dr. Wolfgang Stadler MA from the University of Innsbruck and the Department of Subject Didactics. They helped create an environment that provided the financial security and flexibility to be able to carry out this research.

In addition, I want to extend my sincere thanks to everyone who participated. I would like to thank the students from the University of Innsbruck for their genuine interest and hard work as raters as well as the speakers who volunteered to be filmed for this experiment. Special thanks go to the teachers that successfully encouraged their students to participate in the speaking test simulations: Rosmarie Knoflach, Marietta Heel, Peter Samuda and Michael Schober. I am also particularly thankful to the six expert raters whose ratings were instrumental for this study.

I would like to thank Prof Judit Kormos, Prof Tineke Brunfaut, Dr John Pill, Prof Patrick Rebuschat, Prof J. Charles Alderson as well as the members of the Language Testing Research Group at Lancaster University, who provided encouragement and feedback at various stages. I also want to thank Dr Larry Davis from ETS and Dr Michael Linacre for

their help. My special thanks go to Dr Rita Green who introduced me to statistics many, many years ago and opened up a whole new world to me.

I am extremely grateful to my colleagues at the Language Testing Research Group of the University of Innsbruck: Eva Konrad, Franz Holzknicht, Matthias Zehentner, Elisa Guggenbichler, and Veronika Schwarz. I cannot even begin to express my thanks to Carol Spöttl, whose vision made all of this happen, Sigrid Hauser, whose warm-hearted generosity remains unparalleled, and Benjamin Kremmel, who somehow manages to keep work challenging, fun, interesting and worthwhile on so many levels.

Finally, I would like to thank my parents for providing me with the foundations that allowed me to dream big. I'm also deeply indebted to my wonderful friends in Innsbruck and elsewhere for being my life support system during rough times and making sure I did not forget about the best things in life. Finally, I could not have started, endured and completed this journey without the unwavering support of my best friend and husband, Hans.

CONTENTS

1	Introduction	18
1.1	Background to the Research	18
1.1.1	Rationale.....	18
1.1.2	Problem statement	20
1.1.3	Thesis Aims.....	22
1.1.4	Definitions.....	22
1.2	Outline of the Dissertation	25
2	Literature Review.....	28
2.1	Outline.....	28
2.2	Features of Rater-mediated Assessment	28
2.3	Validating Rater-mediated Assessment	33
2.3.1	Evolution of Validity Theories.....	33
2.3.2	Weir's Socio-cognitive Framework	34
2.3.3	Kane's Argument-based Approach	35
2.3.4	The Argument-based Approach in Rater-mediated Assessments	37
2.4	Defining Rating Quality.....	41
2.4.1	Measures of Rating Quality.....	41
2.4.2	Describing Rating Patterns.....	44
2.5	Rater Cognition.....	47
2.5.1	Rater Attributes	48
2.5.2	The Rating Process.....	56
2.5.3	The Role of Rating Scales.....	65
2.5.4	Summary Rater Cognition.....	69
2.6	Rater Cognition from the Perspective of Information Processing and Decision Making 72	
2.6.1	Key Concepts in JDM	73
2.6.2	Modelling Decision-making Processes	76
2.6.3	Individual Differences in Decision-making Processes.....	81
2.6.4	Judgement and Decision Research in Educational Measurement.....	87
2.7	Variables and Research Questions.....	89
3	Methodology	91
3.1	Mixed Methods in Rater Cognition Research.....	91
3.2	Selected Approach	94
3.3	Research Context	99
3.4	Participants.....	103
3.4.1	Speakers	103
3.4.2	Expert Raters	104
3.4.3	Raters.....	104
3.5	Creating and Selecting Performance Samples	109
3.5.1	Video Recording	109
3.5.2	Reference Scores	110
3.6	Instruments.....	113
3.6.1	Austrian Matura B2 Rating Scale.....	113
3.6.2	Online Rating Platform	116
3.6.3	Cognitive Tests.....	118

3.6.4	Questionnaires and Self-report Data	122
3.7	Procedure	126
3.7.1	Scale Familiarisation Session.....	126
3.7.2	Rating Sessions	128
3.7.3	Cognitive Testing	129
3.8	Analysis.....	130
3.8.1	Rating Data (Study 1).....	131
3.8.2	Thematic Analysis (Study 1).....	137
3.8.3	Time Stamp Data (Study 2).....	138
3.8.4	Cognitive Test Scores (Study 2).....	141
3.8.5	Questionnaire Data (Study 2).....	142
3.8.6	Inspection and Transformation of Variables (Study 2).....	143
3.8.7	Case studies (Study 3).....	144
3.9	Summary	145
4	Results I: Rating Quality and Scale Use	146
4.1	Outline.....	146
4.2	Interrater Reliability (CTT Approach).....	147
4.3	Rater Severity and Consistency	149
4.3.1	Wright Map	149
4.3.2	Group-level Statistics	151
4.3.3	Rater Facet Measurement Results	154
4.3.4	Criterion Measurement Results	156
4.3.5	Global Model Fit	158
4.3.6	Criterion-specific RSM Models	160
4.3.7	Summary Rater Severity	164
4.4	Rater Accuracy (CTT approach).....	165
4.5	Rater Accuracy (MFRM).....	169
4.5.1	Wright Map	169
4.5.2	Group-level Statistics	170
4.5.3	Rater Measurement	171
4.5.4	Criterion Measurement.....	173
4.5.5	Performance Measurement.....	174
4.5.6	Global Model Fit	175
4.5.7	Rater-criterion Interactions	175
4.5.8	Criterion-specific RAM Models.....	180
4.5.9	Summary Rater Accuracy	185
4.6	Perception Data	186
4.6.1	General Confidence in Rating Decisions	186
4.6.2	General Perceived Difficulty of Rating.....	188
4.6.3	Perceived Difficulty of Using the Rating Scale	190
4.6.4	Open-ended Justifications	191
4.6.5	Comparison across criteria	199
4.6.6	Summary Perception Data.....	205
4.7	Summary	206
5	Results II: Exploring the Role of Cognitive and Psychological Attributes in	
	Rater Cognition	208
5.1	Outline.....	208
5.2	Rating behaviour metrics	209
5.2.1	Deliberation Time (DT)	209

5.2.2	Time to First Decision (TTFD)	210
5.2.3	Revision Count (RC)	211
5.2.4	Intercorrelations Between Dependent Variables	212
5.3	Cognitive Attributes	213
5.3.1	Descriptive Statistics	213
5.3.2	Cognitive Attributes and Rating Quality	215
5.3.3	Cognitive Attributes and Rating Behaviour	218
5.4	Preferred Cognitive Processing Mode (REI-40)	220
5.4.1	Descriptive Statistics and Scale Properties	220
5.4.2	Preferred Cognitive Processing Mode and Rating Quality	223
5.4.3	Preferred Cognitive Processing Mode and Rating Behaviour	226
5.5	General Decision-making Style Inventory (GDMSI)	227
5.5.1	Descriptive Statistics and Scale Properties	227
5.5.2	Decision Making Style and Rating Quality	229
5.5.3	DMS and Rating Behaviour	236
5.6	Summary	237
6	Results III: Rater Case Studies	239
6.1	Outline	239
6.2	Selected Cases	239
6.3	Case Study 1: Stormi	241
6.4	Case Study 2: Lexi	251
6.5	Case Study 3: Mary	260
6.6	Case Study 4: Betty	270
6.7	Between-case Comparisons	282
7	Summary and Discussion	284
7.1	Outline	284
7.2	Introduction	284
7.3	Study 1: Rating Quality and Scale Use	286
7.3.1	Summary	286
7.3.2	Severity – Discussion	287
7.3.3	Accuracy – Discussion	290
7.3.4	Perception Data – Discussion	292
7.3.5	Factors Influencing Rating Quality	293
7.4	Study 2: Exploring the Role of Cognitive and Psychological Attributes in Rater Cognition	297
7.4.1	Summary	297
7.4.2	Cognitive Variables – Discussion	298
7.4.3	Preferred Cognitive Processing Mode – Discussion	299
7.4.4	Decision-making Styles – Discussion	301
7.4.5	Rating Behaviour Metrics – Discussion	304
7.5	Study 3: Rater Case Studies	306
7.5.1	Summary	306
7.5.2	Discussion	307
7.6	Theoretical implications	312
7.6.1	Raters	313
7.6.2	Rating Scale	314
7.6.3	Rating Process	318

7.7	Implications for a Validity Argument	323
7.8	Practical Implications.....	325
7.9	Research Method	328
8	Conclusion.....	330
8.1	Limitations	330
8.2	Recommended Further Research	335
8.3	Conclusion	338
9	References	339
	Appendix A Variable Operationalisations	373
	Appendix B Rating Materials.....	375
	Appendix C Speaking Tasks	378
	Appendix D Rating Plan for Reference Scores.....	381
	Appendix E Reference Scores for Selected Speakers	382
	Appendix F Rating Form	384
	Appendix G Intercorrelations Between Measures of Rating Behaviour.....	386
	Appendix H Intercorrelations Between Rater Quality and Rating Behaviour.....	387
	Appendix I Code Book.....	389
	Appendix J Cognitive Tests	392
	Appendix K General Decision-Making Style Inventory (Scott & Bruce, 1995)	401
	Appendix L Rational-Experiential Inventory (Pacini & Epstein, 1999).....	403
	Appendix M Participant Information (for the Speakers)	406
	Appendix N Consent Form and Non-Disclosure Agreement (for the Raters).....	408
	Appendix O Summary of Study 2 Results	409

LIST OF TABLES

Table 2.1 Example warrant and assumptions related to the evaluation inference (adapted from Table 2 in Knoch & Chapelle, 2017, p. 7)	39
Table 3.1 Research design of three linked studies	96
Table 3.2 Overview of raters.....	108
Table 3.3 Rater measurement report	111
Table 3.4 Mean fair average scores in Session 1 and 2	112
Table 3.5 Descriptors for minimum pass level	115
Table 3.6 Overview of cognitive tests in order of presentation	122
Table 3.7 Overview of self-report data collection points in experiment.....	123
Table 4.1 Comparison of most extreme rater pairs in terms of consensus and consistency	148
Table 4.2 Descriptive statistics of inter-rater reliability measures.....	149
Table 4.3 Summary statistic for MFRM analysis of rater severity and fit.....	152
Table 4.4 Measurement report for rater facet.....	154
Table 4.5 Summary of raters (number and percent) within and outside of fit thresholds	156
Table 4.6 Criterion measurement report	158
Table 4.7 Residuals from FACETS analysis (listed by rater).....	159
Table 4.8 Summary statistics of criterion-specific RSM analyses.....	161
Table 4.9 Pearson correlations between criterion-specific severity measures	162
Table 4.10 Number and percentage of misfitting and overfitting raters across criteria	163
Table 4.11 Misfitting and overfitting raters across criteria (based on infit MS).....	163
Table 4.12 Consensus and consistency of participant ratings with expert benchmarks	166
Table 4.13 Cross-tabulation of North's overall scores and the reference scores.....	168
Table 4.14 Summary statistic for rater accuracy analysis.....	171
Table 4.15 RAM estimates and fit statistics for rater facet.....	172
Table 4.16 Accuracy measures and fit statistics for the criterion facet	174
Table 4.17 Accuracy estimates and fit statistics for underfitting speakers	175
Table 4.18 Summary statistics for the exploratory interaction analysis	176
Table 4.19 Significant rater-criterion interactions	179

Table 4.20 Summary statistics criterion-specific RAM analyses	180
Table 4.21 Pearson correlations between criterion-specific accuracy measures	181
Table 4.22 Most accurate raters by criterion.....	183
Table 4.23 Least accurate raters by criterion	184
Table 4.24 Self-reported confidence about rating decisions (percentage row-wise)	187
Table 4.25 Perceived difficulty of rating the performance (percentage row-wise)	189
Table 4.26 Perceived difficulty of applying the four sub-scales.....	191
Table 4.27 Example comments illustrating inhibitors of rating.....	193
Table 4.28 Number and percentage of comments on inhibiting factors	196
Table 4.29 Example comments illustrating facilitators of rating.....	197
Table 4.30 Number and percentage of comments on facilitating factors	199
Table 4.31 Distribution of aspects inhibiting rating decisions across criteria (percentage row-wise)	203
Table 4.32 Distribution of aspects facilitating rating decisions across criteria (percentage row-wise)	204
Table 5.1 Descriptive statistics for mean deliberation time (measured in seconds) for each set of performances and session.....	209
Table 5.2 Descriptive statistics for mean time to first decision (measured in seconds) for each set of performances and each session	211
Table 5.3 Descriptive statistics for revision count across sessions.....	212
Table 5.4 Descriptive statistics for cognitive tasks (N = 39)	214
Table 5.5 Intercorrelations between cognitive test scores	215
Table 5.6 Spearman correlations between rater severity and cognitive tasks.....	216
Table 5.7 Spearman correlations between rater accuracy and cognitive tasks	218
Table 5.8 Spearman correlations between deliberation time (DT) and cognitive tasks	219
Table 5.9 Spearman correlations between time to first decision (TTFD) and cognitive tasks.....	219
Table 5.10 Spearman correlations between number of revisions and cognitive tasks..	220
Table 5.11 Descriptive statistics for Rational Experiential Inventory (REI-40).....	220
Table 5.12 Intercorrelations and reliability coefficients for Rational Experiential Inventory (REI-40).....	221
Table 5.13 Rational and experiential processors based on PSI scores.....	222

Table 5.14 Spearman correlations between severity and preferred cognitive processing mode (N = 39)	224
Table 5.15 Summary of Man-Whitney U tests between rational and experiential processors and rating severity	224
Table 5.16 Spearman correlation between measures of accuracy and preferred cognitive processing mode (N = 39)	225
Table 5.17 Summary of Mann-Whitney U tests between rational and experiential processors and rating accuracy.....	226
Table 5.18 Spearman correlations between measures of rating behaviour and preferred cognitive mode (N = 39)	227
Table 5.19 Descriptive statistics of General Decision-Making Style Inventory (GDMSI)	228
Table 5.20 Spearman intercorrelations and reliabilities for the General Decision-Making Styles Inventory (GDMSI).....	228
Table 5.21 Correlations between measures of severity (MFRM) and decision-making style (DMS).....	230
Table 5.22 Summary of multiple regression analysis for severity (RSL).....	232
Table 5.23 Summary of multiple regression analysis for severity (ASL).....	233
Table 5.24 Correlations between measures of accuracy and decision-making style (DMS)	234
Table 5.25 Summary of multiple regression analysis for accuracy (RSL)	236
Table 5.26 Spearman correlations between rater behaviour metrics and decision-making styles (DMS)	237
Table 6.1 Key selection attributes for case studies	240
Table 6.2 Stormi's comments about the criteria.....	245
Table 6.3 Lexi's comments about the criteria.....	255
Table 6.4 Mary's comments about the criteria.....	264
Table 6.5 Betty's comments about the criteria.....	274
Table 6.6 Detailed profiles of Mary, Betty, Stormi and Lexi	282

LIST OF FIGURES

Figure 2.1 Fulcher's (2003, p. 115) expanded model of speaking test performance	30
Figure 2.2 Eckes' (2015, p. 49) conceptual-psychometric framework.....	32
Figure 2.3 Model of the rating process by Lumley (2005)	59
Figure 2.4 Bejar's (2012) descriptive model of the rating process.....	61
Figure 2.5 Simplified model of rating process (based on Bejar, 2012; McNamara, 1996; Wolfe et al., 1998).....	70
Figure 2.6 Wickens' model of information processing	73
Figure 2.7 Multi-attribute decision models (Newell & Bröder, 2008, p. 200)	78
Figure 3.1 Overview of quantitative and qualitative data integration.....	98
Figure 3.2 Sample individual long turn.....	101
Figure 3.3 Sample paired activity task.....	101
Figure 3.4 Screenshot of rating as presented in Qualtrics.....	117
Figure 3.5 Self-report items after first half of rating session	124
Figure 3.6 Self-report items after second half of rating session	124
Figure 3.7 Self-report items and justifications at the end of both rating sessions	125
Figure 3.8 Screenshot of instructions prior to rating performances.....	129
Figure 4.1 Measurement rulers for RSM analysis	151
Figure 4.2 Measurement rulers for RAM analysis.....	170
Figure 4.3 Bias diagram of interactions between raters and criteria (statistically significant interactions are circled).....	178
Figure 4.4 Confidence in rating decisions.....	188
Figure 4.5 Difficulty of rating.....	189
Figure 4.6 Comparison of mean perceived difficulty for each criterion and session ...	191
Figure 6.1 Stormi's DMS profile.....	243
Figure 6.2 Stormi's notes on P01 (rated 9th in Session 1 with 8 for TA and 7 for FLIN, RSL, and ASL).....	247
Figure 6.3 Stormi's notes on P04 (rated 6th in Session 2)	247
Figure 6.4 Lexi's DMS profile	253
Figure 6.5 Lexi's notes on P01 (rated 15 th in Session 1).....	257
Figure 6.6 Lexi's notes on P04 (rated 3rd in Session 2).....	257

Figure 6.7 Mary's DMS profile	263
Figure 6.8 Mary's notes on P01 (rated 12th in Session 1).....	266
Figure 6.9 Mary's notes on P04 (rated 7th in Session 2).....	266
Figure 6.10 Betty's DMS profile	272
Figure 6.11 Betty's notes on P01 (rated 15th in Session 1).....	277
Figure 6.12 Betty's notes on P04 (rated 2nd in Session 2).....	277
Figure 7.1 Expanded model of rater cognition in rating speaking.....	321

LIST OF ABBREVIATIONS AND ACRONYMS

ASL	Accuracy of Spoken Language (criterion in Austrian rating scale)
Avg	Average
CEFR	Common European Framework of Reference
CEST	Cognitive-experiential self-theory
CTT	Classical Test Theory
EA	Experiential ability (REI-40)
DMS	Decision Making Style
DT	Deliberation Time (rater behaviour metric)
EE	Experiential engagement (REI-40)
EFL	English as a Foreign Language
EP	Experiential preference (REI-40)
FLIN	Fluency and interaction (criterion in Austrian rating scale)
GDMSI	General Decision-Making Style Inventory
IRT	Item Response Theory
L1	Primary or first language(s)
L2	Second or non-primary language(s)
M	Mean
MDN	Median
MS	Mean square
MS _w	Weighted mean square, also infit
MS _u	Unweighted mean square, also outfit
NNS	Non-native speaker
NS	Native speaker
Obs.	Observed
PSI	Processing Style Inventory (REI-40)
RAM	Rater Accuracy Model
RA	Rational ability (REI-40)
RC	Revision Count (rater behaviour metric)
REI-40	Rational Experiential Inventory (40-item version)
RMSE	Root mean-square measurement error
RP	Rational preference (REI-40)
RQ	Research question
RSM	Rating Scale Model
RSL	Range of spoken language (criterion in Austrian rating scale)
SD	Standard deviation
SE	Standard error
SLA	Second Language Acquisition
SPSS	Statistical Package for the Social Sciences
TTFD	Time to first decision (rater behaviour metric)
TA	Task achievement (criterion in Austrian rating scale)
t _w	Standardized infit
t _u	Standardized outfit

1 Introduction

1.1 Background to the Research

1.1.1 Rationale

During foreign language performance tests, candidates engage with language tasks and produce spoken or written language, referred to as the *performance*. Human raters observe and judge the performance “using an agreed *judging process*” to attain a score (McNamara, 1996, p. 10, emphasis in the original). Using raters to judge language proficiency, however, adds a potentially influential factor to the measurement process (Eckes, 2015; McNamara, 1996), particularly as rating is itself a highly complex cognitive process (Alderson et al., 1995). The examinees are given a task that is believed to sample the underlying construct and the rater observes the resulting language performance, forms a representation of this performance, and compares it with the more or less explicitly communicated rating criteria (Bejar, 2012; Eckes, 2015; Lumley, 2005; McNamara, 1996; Wolfe, 1997, 2006; Wolfe et al., 1998). Thus, the relationship between candidate performance and resulting score is not straightforward (Eckes, 2015), and has been found to be influenced by numerous factors (Cumming, 1990; Lumley, 2002, 2005). Ultimately, the use of human judgment constitutes a potential source of construct-irrelevant variance (Bachman & Palmer, 1996) and poses a threat to the validity of an examination’s scores (Kane, 2013; Knoch & Chapelle, 2017).

In response to these issues concerning rating, the emerging field of rater cognition research investigates how certain rater attributes (e.g., L1, accent familiarity, experience) may impact scoring patterns and the mental processes involved when

raters allocate scores (e.g., which criteria are heeded, how the scales are being used, typical steps of the rating process) (Bejar, 2012). This research has contributed to a better understanding of how certain attributes may affect rater consistency patterns as well as mental processes. As a result, test providers may opt to document certain rater attributes that are considered salient to rater cognition and recruit raters accordingly. However, studies in the areas of rater training and experience show that rating is a highly complex process and even experienced and trained raters produce erratic ratings (Eckes, 2015; Lim, 2011), suggesting that there may be other features related to rating – rater characteristics, aspects of the rating scale as well as interactions between these two – that still need to be explored.

As will be argued in this dissertation, rater cognition in relation to assessing speaking is still under-researched. This is problematic for a range of reasons. First, there has been an increased focus on assessing communicative skills (Fulcher et al., 2011) and performance assessment in commercial assessment settings as well as language classrooms, leading to a “continued reliance on human markers” (Leighton, 2012, p. 1). Achievements in digital technologies have made it possible to automatically assess highly-controlled samples of language for some purposes, but testing speaking in performance-based assessment necessitates the use of human judgement and is likely to do so in the future as well – particularly in the context of tests with smaller populations and in educational settings. Second, the increased focus on speaking skills also increases the need for developing rating scales and training raters. Rater training has been found to improve intra-rater consistency (e.g., Kim, 2015), but appears to fail to impact equally all features of rating quality (Davis, 2008; Weigle, 1998; Kang et al., 2019), and lead to predictable rating patterns for all raters (Eckes, 2015; Lim,

2011). Thus, there is a need to investigate more closely the potential impact of rater attributes on rating patterns.

1.1.2 Problem statement

So far, language testing research has increasingly investigated rater cognition to address the issues outlined above, but studies that examine the processes involved in the rating of speaking as opposed to writing and the use of analytic rating scales remain scarce (Eckes, 2005; Yan, 2014; Zhang & Elder, 2011). The great bulk of research on rater cognition in speaking is set in the context of highly standardized tests where raters are selected, trained, and monitored (e.g., Yan, 2014), and tend to use holistic scales (e.g., Cumming, 1990; Davis, 2015). While the effects of test taker attributes like physical, psychological and perception characteristics have been at the centre of numerous cognitive validation studies, there is a need to focus more on the psychological and cognitive dimensions of rater cognition, particularly in the context of speaking. There is also a need to consider rater cognition, more generally, in other environments beyond that of large-scale international, standardised tests.

This research project was conducted in the context of the Austrian school-leaving examination, also called the *Matura*. This high-stakes test marks the end of upper-secondary school education and entitles graduates to enrol in tertiary education programmes at universities or colleges (though several programmes such as medicine and psychology may require additional entry tests). Between 2007 and 2015, at a time when many European curricula and school-leaving examinations were being reformed (e.g., Alderson, 2011; Brunfaut & Harding, 2018), the foreign languages Matura also underwent a major overhaul. In its original form, the examination was set, administered, and graded by the class teacher and would often hinge on the assessment

of content knowledge related to history, cultural studies, or literature as well as practical language use. The inclusion of the competence levels described in the Common European Framework of Reference (CEFR; Council of Europe, 2001) into the national curricula for foreign languages paved the way for standardizing the exit examination and shifting classroom teaching towards a skills- and competence-oriented language pedagogy. One consequence of the reform was the obligatory use of analytic assessment scales for the final writing and speaking examinations (Spöttl et al., 2016) which were developed by a project coordinated at the University of Innsbruck.

The scales heralded a paradigm shift in Austrian classroom assessment practices and were designed to serve multiple purposes. For one, the speaking and writing scales published by the Ministry were the first of their kind specifically designed for the Austrian school context. To ease teachers into using the speaking scales, they were deliberately created to mirror several key features of the Austrian writing scales which had been released earlier. These features included 1) four criteria, 2) eleven bands with a maximum score of ten, 3) separate criteria for linguistic range and accuracy, and 4) a criterion for task achievement (Holzknecht et al., 2018). Secondly, the analytic nature of the scale offered the opportunity of more fine-grained feedback, which may be argued to have particular merit for an assessment that is embedded within an educational system. Thus, the rating scale not only serves the function of an assessment tool during the test, but its potential influence extends beyond the day of the examination. While the scales contributed to fostering a shared understanding among Austrian teachers of what constituted spoken communication in a foreign language, there were and still are limited training opportunities for teachers in using

the scale. This raises the question as to how soon-to-be-qualified and inexperienced teachers might be coping with the task of using an analytical rating scale.

1.1.3 Thesis Aims

This dissertation outlines a research project that sought to investigate the rating quality of a group of novice raters, and, drawing on concepts from the interdisciplinary field of judgement and decision-making research (JDM), explored the contribution of cognitive and psychological attributes on the consistency of the novice raters' scoring decisions and observable rating behaviours. The focus of this study will be on dimensions of rating quality and observable rating behaviours, and how they might be impacted by previously under-researched rater characteristics.

Set in the context of a recently reformed national school-leaving examination the study examined how individual raters handle the demands of assessing speaking performances with an analytic rating scale. The study recruited pre-service teacher students who were novice to rating for several reasons: 1) novices have not yet developed the coping mechanisms or strategies that help more seasoned raters with their task; 2) without mediating strategies and through a more controlled sample, the effects of cognitive and psychological attributes as well as challenges may become more visible; and finally 3) as Fulcher (2003) suggested, observing how raters apply a rating scale without intensive training provides clearer insights into the strengths and weaknesses of the assessment instrument.

1.1.4 Definitions

As the assessment of language ability takes place in various contexts and for various purposes, there are considerable differences in *who* is charged with assessing the

candidates' performances and *how* the assessment is carried out. Accordingly, a range of terms evolved around rater-mediated assessment. A few key terms which will be used throughout this dissertation will be defined below.

Raters. The term *rater* will be used to refer to a person who judges the quality of a language performance based on a set of criteria. The term is understood to be synonymous with *scorer* or *marker*. Depending on the context of the assessment, raters can range from being highly trained individuals who work full or part time for a language testing institution to teachers who seasonally return to the task of rating performances as the school year progresses. For the purposes of this dissertation, the term *rater* is used to refer to individuals who watch a speaking performance and allocate a score.

Rating scale. The term *rating scale* will be used to refer to documents that are used for the purpose of assessing language performances. The term is considered a synonym for (*scoring*) *rubric*. Rating scales are usually either *holistic* or *analytic*. As Harsch and Martin (2013) discuss, there are some discrepancies as to how the terminology as first defined by Hamp-Lyons (1991) is employed across studies. In line with their argument, the term *holistic* will be understood to refer to rating methods producing a single score for a performance and *analytic* will be used to refer to rating multiple traits of a performance described in a scale with multiple *criteria* (or, dimensions) and producing multiple scores, one for each dimension. The rating scale used in the current study contained four separate criteria and ten distinct rating *bands* or levels of ability.

Rater cognition. For this thesis, *rater cognition* will be understood to encompass the various mental activities which are taking place when raters allocate scores to a

language performance. This process is generally understood to be highly complex and shaped by various factors (e.g., Eckes, 2015; Lumley, 2005). To be able to decide about a speaker's ability, raters perceive and process information as the performance unfolds, notice certain language features, compare their perception (also called mental representation) of the performance with internal or external criteria, and assess or weigh the various features of a performance in light of the criteria (but also in consideration of the context in which the assessment takes place). The various cognitive processes involved in rating language ability can be conscious to the raters, but also subconscious. Therefore, rater cognition research is concerned with uncovering the effects of rater attributes on rating patterns and the mental activities that take place when raters allocate scores (Bejar, 2012).

Rating behaviour. When raters allocate scores to candidate performances, they also display a range of observable behaviours. Building on the process tracing approach to investigate cognition, such observable behaviours can be viewed to be indirect concomitants of the cognitive processes taking place (Schulte-Mecklenbeck et al., 2011). The behaviours observed in the context of this dissertation were mainly based on time stamps collected during the rating in an online rating form, the revisions of rating decisions as well as the notes that raters took during the rating sessions.

Rating quality. As will be discussed at some length, there is not one distinct index or measure to determine the rating quality of any given rater. Instead, it is the context of the test or assessment which determines which rating decisions can be considered *good* or *useful*. In some contexts, the emphasis may be that raters are interchangeable while other contexts acknowledge the individual expertise each rater brings to the task of rating and that some disagreement is to be expected (e.g., McNamara, 1996). In this

dissertation, rating quality is understood as a multi-faceted phenomenon which cannot be adequately captured by a singular metric (Harsch & Martin, 2013). Rating quality is determined by the extent to which individual raters agree with each other (inter-rater reliability), with the reference scores (accuracy) and with their own ratings (intra-rater reliability or consistency).

1.2 Outline of the Dissertation

Chapter 2 examines the specific features of rater-mediated assessment and the theoretical issues of assessing foreign language speaking ability. This includes an examination of how validation theory is applied to this specific testing setting and how rater quality can be defined and measured. Previous research will be reviewed to determine how rater cognition is shaped by attributes such as languages spoken, experience and expertise. Furthermore, the chapter will discuss the extent to which the rating process might be shaped by the modality of the performance (spoken vs. written) and rating scale type (analytic vs. holistic). Based on this review, an argument will be developed for investigating rater cognition, and the role of cognitive and psychological attributes in the context of speaking examinations, more closely than is typically theorized in speaking assessment models. Key concepts from psychology as well as judgement and decision research will be introduced briefly before discussing a small set of language testing studies which do incorporate aspects of applied psychology that may be pertinent to rater cognition.

Chapter 3 outlines the mixed-methods research design and methodology of this study. This includes a description of the research context, the various groups of participants, and the development of a set of spoken performances which were specifically created

for this purpose. Next, the chapter provides details on the cognitive test battery, the questionnaires, and the online rating environment. Procedures of rater training, the rating sessions and the cognitive testing are summarized before elaborating on data analysis techniques.

Chapter 4 reports on the results of a first study which investigated the warrants connected to the evaluation inference and, more specifically, to the claim that raters rate reliably when using the Austrian assessment scale. The analysis investigated rating quality, and, more specifically, rater severity, rater fit, bias and accuracy as well as a more qualitative investigation of how confident and comfortable raters were with using the rating scale.

Chapter 5 summarizes and presents the results of an exploratory study which sought to apply concepts from judgment and decision-making research to rater cognition research. The study investigated whether there were associations between rating quality and specific aspects of rater behaviour on the one hand and rater attributes such as cognitive ability, preferred modes of processing information and decision-making styles on the other.

Chapter 6 presents a set of four case studies selected on an extreme cases approach. Data from the previous two chapters are collated and broken down to focus on four extreme cases: two very accurate and two inaccurate raters. Similarities as well as differences between the raters and how each appears to approach the task of rating spoken performances will be investigated and presented in detail.

Chapter 7 summarizes the key findings for each of the three studies and discusses them in relation to findings in previous studies. It will then provide a broader discussion of

the issues encountered and raised, and the various implications of this research on theory, practice, and methodology in the area of language testing and assessment.

Chapter 8 provides an overview of the approach taken, methodology and key issues identified in the three studies. Limitations of this study and potential areas of future research will be outlined.

2 Literature Review

2.1 Outline

Section 2.2 will first define rater-mediated measurement within the context of speaking examinations and the most common conceptualizations of the components of speaking examinations. This includes discussing the various factors that may contribute to the outcome of a speaking test. Inconsistent rating is a core concern of language assessment as the reliability of the rating process also threatens the validity argument of a given language test. Some of the most influential validation theories in relation to rater-mediated assessment will be discussed briefly in Section 2.3 to clarify how reliability is connected to validity concerns. Next, Section 2.4 will define current understanding of what constitutes rating quality and descriptions of rater effects. Section 2.5 will then focus on rater cognition as a specific area of language testing research and identify key issues that are typically investigated. To explore the role of rater attributes and rater cognition in the context of speaking from a perspective that incorporates considerations of the cognitive processing preferences and capacities of raters, Section 2.6 introduces key concepts from cognitive and economic psychology. Finally, Section 2.7 presents the research questions and the variables included in this study.

2.2 Features of Rater-mediated Assessment

Language teaching and the assessment of language learning has increasingly shifted towards emphasizing communicative skills. However, the testing of productive second language skills, and speaking skills in particular, poses a great challenge for any test

developer in terms of practical concerns and potential threats to reliability (Alderson et al., 1995; Csépes & Együd, 2005; McNamara, 1996; Underhill, 1987). This situation has led to reservations among some test developing bodies to assess speaking skills at all (Fulcher, 2003). Nonetheless, there has been an increased and sustained effort to develop and improve the testing of speaking skills in performance assessment contexts (Bachman, 2000; Fulcher, 2003; McNamara, 1996; Taylor & Galaczi, 2011).

In its broadest sense, “measurement . . . is defined as the assignment of numerals to objects or events according to rules” (Stevens, 1946, p. 677). The “rules” of how or when to allocate which scores are operationalized via rating scales (or rubrics) (Fulcher, 2003). Thus, rating a language performance constitutes forming “judgments of quality against some rating scale” (McNamara, 1996, p. 3), which is anything but a trivial task. As Myford and Wolfe (2003) emphasize, raters do not just record an outcome of the test; “rather, their ratings are rooted in observation, interpretation, and, perhaps most importantly, the exercise of personal and professional judgment” (p. 389). One way of framing the problem is to view rating as a problematic, “complex and error-prone cognitive process” (Cronbach, 1990, as cited in Myford & Wolfe, 2003, p. 392). McNamara (1996, p. 117), on the other hand, argues that judgements about performance are bound to be complex and nuanced and that it is this quality that makes them worthwhile in the first place.

Arguably the greatest obstacle that language testers must face when assessing productive language skills is to deal with unwanted score variability caused by raters. If test scores are to be informative and useful to stake-holders, any score variance should be directly attributable to candidate ability. Quantitative and qualitative investigations into test scores and rater cognition, however, have repeatedly shown

that other factors contribute to test outcomes (e.g., Eckes, 2015; Lumley, 2005). To guide research into rater-mediated language tests, several models were developed to capture the various factors involved in assessing speaking. Fulcher (2003), for example, built on previous work by McNamara (1996) and Skehan (2001) and created a componential model (see Figure 2.1). According to Fulcher (2003, p. 113), the three main contributions to a score are the test taker ability, the difficulty of the task and the conditions of the assessment. The test taker comes to the speaking assessment with certain characteristics, abilities, capacities and knowledge and their performance then depends on the nature of the task and the conditions under which they are assessed.

Figure 2.1 Fulcher's (2003, p. 115) expanded model of speaking test performance



According to Fulcher (2003), it is not “accidental” (p. 115) that the definition of the construct and its operationalisation through the rating scales form a central component in this model. The rating scales shape which features are sought in the performances of the test takers and what kind of inferences can be made about test taker ability. The model also suggests how consistency in applying the rating scale may affect the inferences that can be made about the test takers and the construct. Interestingly, the model considers the various aspects which may impact test takers’ cognition such as their task specific knowledge, capacity for real-time processing, language abilities as defined in the construct and individual variables such as personality. The raters’ performance, on the other hand, is seen to be affected by training and rater characteristics. However, the model offers no detail on the range of rater characteristics that might play a role in rater cognition.

Eckes (2015) takes a different approach that is useful for modelling rater-mediated assessment from a psychometric perspective (see Figure 2.2). His framework differentiates between *distal* and *proximal* facets that affect the ratings or observed scores. According to Eckes (2015), the examinee’s proficiency is the single most important facet and should have the greatest impact on the performance and outcome. An examinee’s proficiency also interacts with the difficulty of the task as a more difficult task might elicit a lower score than a less difficult task. The other three proximal facets are the severity of the rater, the difficulty of the criterion and the scale structure, all of which are not related to the construct but may introduce sources of score variation.

Figure 2.2 Eckes' (2015, p. 49) conceptual-psychometric framework



The distal facets Eckes (2015) presents in the model are the respective features of the examinees, raters and rating situation which may again interact with each other and with the proximal facets. As the author points out, the influence of distal facets may be less direct and diffuse, and they may also interact with one another and the proximal facets in various complex ways. This leads Eckes to conclude that the link between a performance and the observed score is

fragile . . . (as) the score a rater awards to an examinee is the result of a complex interplay between bottom-up, performance-driven processes (e.g., distinct features of the performance) and top-down, theory-driven (knowledge-driven) processes (e.g., expectations based on knowledge of the prior examinee performance or based on gender, age, ethnic, or other social categories). (2015, p. 51)

It is important to note that while Fulcher's model only considers how rater training and characteristics impact the examinees' performances, Eckes' psychometric model includes aspects such as rating context and rater workload. It is also more specific as

far as the attributes of raters as well as examinees are concerned, including the rater's gender, experience, education, and attitudes.

As these two models illustrate, the various factors involved in rater-mediated assessment interact dynamically and shape the outcome of the assessment process. How each component and its interactions with other components affects the score may be difficult to uncover and investigate. Nonetheless, as will be argued in the following section, “worrying about rating” (Hamp-Lyons, 2007) is essential to delivering assessments that are worthwhile.

2.3 Validating Rater-mediated Assessment

2.3.1 Evolution of Validity Theories

Validity is a central concern in second language testing. Along with the purposes and methods of language assessment, the conceptualization of validity has evolved (e.g., Chapelle, 1999). Naturally, the field has brought forth a range of various different stances towards validation (Fulcher, 2015) and continues to discuss further avenues. However, there was agreement early on that validity cannot be viewed as a property or characteristic of a test and that it cannot be established through a single source of evidence or expressed by a single measure (Bachman, 1990; Sireci, 2016; Weir, 2005).

An early theoretical approach distinguished between four types of validity: predictive, concurrent, content and construct validity (e.g., Cronbach & Meehl, 1955). Validation studies would usually be carried out in the form of correlational studies, investigating the relationship of test items to the test construct through factor analysis or comparing scores on similar tests. While the rating of language performance and the possible

effects of the rating process on scores were considered a central matter in test design and implementation, these issues were regarded as a question of test reliability and less immediately concerned with test validity.

Throughout the 1980s and the 1990s, new perspectives on validity theory and test validation emerged, culminating in the revision of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999), Messick's seminal chapter in the third edition of *Educational Measurement* (Messick, 1989), and Bachman's chapter on validation (Bachman, 1990). Bachman provided one of the first translations of this "new" approach into the language assessment domain. Instead of segregating different types, validity came to be viewed as a unitary yet multifaceted construct (Messick, 1989) that can be investigated through various different methods and types of evidence (Bachman, 1990; see also Alderson et al., 1995). Thus, the focus of validation studies shifted towards investigating the validity of score interpretations and use (Xi, 2008). While expanding the scope of validity, this approach posed new methodological challenges for those directly involved in test development. Bachman (1990) observed the "continuous" nature of validation and that "evidence for the validity of test interpretations is of several types, and can be gathered in a number of ways" (p. 289; also Alderson et al., 1995, p. 171).

2.3.2 Weir's Socio-cognitive Framework

The socio-cognitive framework first introduced by Weir (2005) is one of several models that emerged as a reaction to the practical challenges of Messick's (1989) unitary model. Within this framework, Weir distinguishes between context validity, theory-based validity, scoring validity, consequential validity and criterion-related

validity. Thus, the socio-cognitive model returns to a more compartmentalized view of validity which resembles earlier theories more than Messick's unified concept. Within this framework, establishing scoring validity is defined as an *a posteriori* analysis of potential validity evidence on the extent of rater agreement, consistency, severity and intra-rater reliability (Weir, 2005, p. 22-40).

The strength but also weakness of Weir's socio-cognitive framework is that it provides a roadmap or extended checklist that is designed to address every critical feature a given test may have. Besides not further progressing validity theory (Fulcher, 2015), one of the most significant criticisms of this approach is that such a technical view may lead to gathering a plethora of evidence without showing how this evidence can be integrated or combined into a coherent argument (Chapelle et al, 2010a). The challenge also lies in defining the scope of validation efforts and a possible saturation point (Kane, 2013) in order to avoid making test validation a potentially "never-ending process" (Anastasi, 1986, p. 4). Furthermore, there is a danger of amassing confirmatory evidence that is conveniently available (Fulcher, 2015) or yielding to what Cronbach labelled "dragnet empiricism": gathering a lot of evidence without testing a clearly formed hypothesis (Cronbach, 1988; 1989). Finally, it is argued that without a clearly defined argument steering the validation effort, the socio-cognitive approach offers little guidance in how to prioritize the research agenda and ensure that enough attention is dedicated toward investigating weaker and critical aspects of the validity argument (Chapelle et al., 2010b; Fulcher, 2015).

2.3.3 Kane's Argument-based Approach

Argument-based validation (Kane, 1992, 2013; Shepard, 1993) has emerged as a popular approach within the language testing community (Fulcher, 2015). The focus

lies on building a clear, coherent and plausible argument towards justifying the interpretations and uses of the assessment under investigation (Kane, 1992, 2013). In doing so, the argument-based approach sidesteps some of challenges associated with Messick's unified model of validity without regressing back to conceptualizing validity as separate types of validities (e.g., Fulcher, 2015, when discussing Weir, 2005). Looking at validity as an argument offers a practical as well as rigorous strategy to test validation (Chapelle et al., 2010b; Kane, 2013; Knoch & Chapelle, 2017; Sireci, 2016) in that it can be based on a formal argument without requiring "formal theories" (Kane, 1992, p. 534). This is a clear advantage given that attempting to define test constructs remains a recurring and unresolved challenge in language assessment (Chapelle, 1999). Owing to its theoretical robustness and applicability, the principles of the argument-based validation framework have been embedded into the AERA/APA/NCME *Standards* since the 1999 edition (Davidson, 2000; Sireci, 2016).

At the heart of the argument-based approach lies the intent to build a plausible and convincing case that the intended interpretation(s) or use(s) of test scores are indeed valid, an *interpretive argument*. Constructing an interpretive argument rests on identifying the areas for which evidence will be needed and integrating this evidence into a "well-grounded or firmly backed claim" (Toulmin, 2003, p. 8) that connects inferences, warrants and assumptions beyond test score interpretation and use (Kane, 1992; 2013). Thus, an interpretive argument makes explicit the various aspects of a test score about which inferences are made and the warrants and assumptions that help back each of these inferences.

The meaning of a test score is shaped by various aspects based on decisions taken during test development and implementation. In Kane's original conceptualization he

focussed the model on three such meanings or inferences: *observation*, *generalization* and *extrapolation* (Kane, 1992). The *observation* inference (later also called *scoring* inference in Kane et al., 1999; or *evaluation* inference in Xi, 2008) is based on the assumption that any resulting score has been determined following consistent procedures in terms of test administration and marking. The second inference in Kane's original model is the *generalization* inference. This inference hinges on the claim that any score is equally representative of a candidate's ability despite smaller changes that may occur with respect of the conditions of measurement (i.e., test form, examiner, test centre, etc.) (Kane, 1992). The third inference described by Kane is the *extrapolation* inference. The meaning of a test score rests on the assumption that any score will provide valuable and reliable information on a participant's skill regarding the test construct. So, the meaning of test scores depends on there being some kind of direct relationship between the behaviour observed in a test situation and the behaviour in a non-test situation (Kane, 1992).

Since its first conceptualisation (Kane, 1992), the argument-based model has undergone various revisions by Kane himself (Kane et al., 1999; Kane, 2004, 2006), but also others from within the language assessment field (Knoch & Chapelle, 2017; Xi, 2008). Later iterations reflect the tendency to put a stronger emphasis on test score use and interpretation. Thus, inference categories such as *utilization* (Bachman, 2005 when summarizing Kane, 2004), *decision (making)*, *consequence* or *representation* can be found in later versions (Knoch & Chapelle, 2017).

2.3.4 The Argument-based Approach in Rater-mediated Assessments

Kane's work provides a baseline for developing an argument-based validation programme within language assessment contexts (Chapelle et al., 2008, 2010b;

Shepard, 1993; Xi, 2008). However, as Knoch and Chapelle (2017) highlight convincingly, the therein provided warrants and claims emphasize the testing of receptive skills and are not readily useful and sufficiently specific for framing the validation agenda of a performance-based test. Furthermore, the general warrants tend to define the significance of the rating process more narrowly and are mainly concerned with reliability. As Knoch and Chapelle (2017) argue, language testers investigating rater cognition may look at very diverse phenomena such as the psychometric properties of rating scales, how the scales are linked to the construct(s) or rater engagement. Drawing on various versions of Kane's framework (2001, 2006, 2013) and Chapelle et al. (2008), Knoch and Chapelle (2017) address this perceived gap by supplementing warrants and assumptions that are specific to performance-based assessment. They demonstrate how rating-related issues penetrate *all* inferences made in an interpretative argument, from *evaluation* to *consequence*. Furthermore, they link each assumption to sources of evidence and methods of data collection and analysis.

For example, the *evaluation* inference is linked to the claim that rating produces scores "with intended characteristics" (Knoch & Chapelle, 2017, p. 7). This claim partly builds on the warrant that "raters rate reliably at task level" (p. 7) which in turn can be established by investigating ten assumptions (see Table 2.1). As this example illustrates, the assumptions and the evidence that would support them reach beyond establishing rater consistency and reliability. Issues related to rater cognition, as in the scoring patterns awarded by raters and rater attributes which may impact on these patterns are intertwined with these assumptions as they relate to individual rating processes, questions of training and support, as well as bias.

*Table 2.1 Example warrant and assumptions related to the evaluation inference
(adapted from Table 2 in Knoch & Chapelle, 2017, p. 7)*

<p>Inference: Evaluation</p> <p>Associated claim: Observations are evaluated using procedures that provide observed scores with intended characteristics</p> <p>Warrant B: Raters rate reliably on task level</p>
<ol style="list-style-type: none"> 1. Raters are able to identify differences in performances across score levels. 2. Raters can consistently apply the scale to test tasks. 3. Raters are comfortable when applying descriptors and confident in their decisions. 4. Raters are thoroughly and regularly trained in use of the scale and sub-scales (if applicable). 5. Sufficient rater support documents with scale exemplifications are available. 6. Raters are suitably qualified. 7. Rating sessions are designed to optimize rater performance. 8. Detectable rater characteristics do not introduce systematic construct-irrelevant variance into task ratings above acceptable levels set by the test designer. 9. The level of rater bias towards particular sub-scales (if applicable) is within acceptable levels set by the test developer. 10. The level of bias raters display against task types or other systematic aspects of the test situation (beyond scale criteria) is within acceptable limits set by the test developer.

Rater cognition is also referred to explicitly when the authors discuss the *explanation* inference (Knoch & Chapelle, 2017). The explanation inference rests on the claim that

scores can be explained by an underlying model of language proficiency and a clearly defined theoretical construct. Here, one assumption that can be explored is whether “raters’ cognitive processes are consistent with the theoretical model of proficiency and/or development” (p.13). It is possible that raters interpret criteria differently to the intentions of the scale developers (e.g., McNamara, 1996) thus undermining the meaning of scores and threatening the validity of the validity argument. Collecting verbal reports during the rating process as suggested by the authors may be one way to investigate this threat.

One strength of the argument-based approach, and in particular an iteration that is tailored to rater-mediated assessment such as the one proposed by Knoch and Chapelle (2017), is that an investigation into the rating process and its outcomes is firmly embedded in the validation process. This is not to say that rating and concerns about rater reliability were regarded as secondary in previous validity models. Weir (2005), for example, grants rating great importance by installing scoring validity next to traditional validities in his socio-cognitive framework. However, the argument-based approach acknowledges how the rating process is linked to some of the most central assumptions that language testers and users make about the meaning of test scores. An argument-based view of validity challenges the field to formulate rater cognition-related propositions more broadly, and to integrate them in the interpretive chain (e.g., Myford, 2012 in reference to Kane, 2006).

2.4 Defining Rating Quality

2.4.1 Measures of Rating Quality

Research into rater effects and rater cognition has led to the development of a vast and potentially confusing array of measures that can be used to investigate the quality of ratings from an individual rater or from rater cohorts (e.g., Bachman; 2004; Myford & Wolfe, 2003). Even the deceptively simple concept of inter-rater reliability has no single definition that would hold across all its applications (Gwet, 2014). According to Weir (2005), research into scoring validity needs to determine “to what extent (raters) are: in overall agreement, ranking a group of students in the same order, rating individuals at the same level of severity, [and] consistent in (their) own judgements during the whole marking process” (p. 34). However, Weir does not specify which statistical procedures or indices would suffice to provide information on each of these four points and depending on the context of the assessment, there might not be one simple and straightforward answer to this question. From the current point of view, there are two clearly distinguishable research paradigms, or measurement models, within which rater performance is conceptualized (Eckes, 2009, 2015; Engelhard & Wind, 2018; Myford & Wolfe, 2003; Green, 2013). These two major traditions are referred to as 1) the observed ratings tradition, and 2) the scaled rating tradition (Wind & Peterson, 2017).

In the context of the observed ratings tradition, any observed score is understood as a combination of a candidate’s true score and measurement error. The emphasis lies on the observed score and how it may or may not relate to the true score of a candidate. Hence, this approach is also referred to as the *True-Score Theory* (McNamara, 1996)

or *Test-Score Tradition* (Wind & Peterson, 2017). In rater-mediated assessment the performance of a rater can be evaluated by comparing the actual score given by an ‘ordinary’ examiner to the idealized score given by a ‘perfect’ examiner (Bachman, 1990; Lumley, 2005). Differences in the score outcomes are considered to be caused by varying degrees of either, severity or error (Wind & Peterson, 2017). There are several measurement models within the observed ratings tradition (e.g., classical test theory, analysis of variance, or factor analysis), which share common approaches such as summing or averaging scores across raters, test components or candidates.

Within the observed ratings paradigm, there are two basic classes of indices that define rating quality. Estimates of *consistency* are calculated by correlating sets of ratings (either by the same rater or different raters) on the same performance and estimates of *consensus* are estimates based on variances (Bachman, 1990; 2004; Brown, 2005). According to Eckes (2015), rater consensus reflects how much raters provide the same rating for the same performance and can either be measured in terms of *exact agreement* which constitutes the proportion of exact matches between sets of rating scores, or even more reliably, through a *chance-corrected agreement coefficient* (e.g., Cohen’s kappa, Gwet’s AC1), where the observed agreement is adjusted by using an estimate of the expected chance agreement (Gwet, 2014; Upton & Cook, 2014). Estimates of interrater consistency, on the other hand, are based on correlations and reflect the extent to which two raters “provide the same relative (...) ranking of the examinees” (Eckes, 2015, p. 42). Two commonly used consistency measures are the Pearson’s product-moment correlation coefficient and the Spearman rank correlation. Pearson’s r is a measure of association that represents the degree of linear relationship between two sets of ratings and is calculated by dividing the covariance of the ratings of two raters by the product of their standard deviations (Colman, 2015). Spearman

rank correlations are a nonparametric alternative to r and measure the monotonic association between two sets of ranks. Another measure particularly useful for ordinal data is Kendall's tau-b which involves calculating the difference between concordant and discordant pairs of rating decisions divided by the total number of pairs (Colman, 2015). Kendall's tau-b is generally found to be more robust than the Spearman rank correlation (Croux & Dehon, 2010) and is more useful if the data is expected to include many tied ranks, i.e., same decisions between two raters (Field, 2014, p. 276).

While measures of consensus or consistency provide a useful first impression of rating data, the currently dominant approach in second language assessment research when investigating rater behaviour is to move away from the observed ratings tradition or at least to supplement these findings with analyses based on the scaled ratings tradition (Wind & Peterson, 2017). Applications of the scaled ratings approach include Rasch measurement theory and item response theory (IRT). Instead of using raw scores, the raw scores are transformed into probabilistic estimates of a student's ability and item difficulty. Students and items are mapped onto a scale measuring a latent ability variable which makes it possible to compare item difficulty and student ability across different populations and items. This solves the great measurement issue of using raw scores because estimates based on Rasch models can be considered stable (or *invariant*) while raw scores are always variable, depending on the ability of the candidates included in the sample and how they interact with the difficulty of the items included in the test. The estimates resulting from Rasch-based analyses allow generalizations about the difficulty of items and ability of candidates beyond a particular sample and predict the likely difficulty of items for an entire population of test takers (Stone & Wright, 1979; see also Bachman, 2004; Eckes, 2015; McNamara, 1996).

Rater-mediated assessment is most commonly investigated via many-facet Rasch measurement (MFRM), which is based on Rasch measurement theory and item response theory (Wind & Peterson, 2017), but extends the dichotomous conceptualisation of these models by at least one more component of the assessment situation. With the observed scores as fixed observations, estimates of the other components, or *facets*, included in the model are transformed and mapped onto a common logit scale (Engelhard & Wind, 2018). As a result, the impact of any facet that is believed to affect candidate scores (e.g., raters, test form, test session, or criteria), can be included in the model and investigated within one frame of reference (e.g., Eckes, 2015; McNamara, 1996).

2.4.2 Describing Rating Patterns

Cronbach described rating as a “complex and error-prone cognitive process” (Cronbach, 1990 as cited in Myford & Wolfe, 2003, p. 391). Consequently, score variability that is not due to variability in candidate ability is a frequently occurring and undesirable consequence of involving raters in the assessment of language ability. Rater cognition research has established that raters vary in systematic patterns which may be due to conscious and unconscious factors influencing the rating process (Bachman et al., 1995; Myford & Wolfe, 2000; Wolfe & McVay, 2012; Eckes, 2012; Wigglesworth, 1993). As a result of this work, various types of rating patterns or rater effects have been identified in language testing contexts (Eckes, 2009; McNamara, 1996; Myford & Wolfe, 2003; Wolfe, 2004). The most frequently discussed effects are:

- leniency/severity
- bias
- halo

- differences in rating scale use
- differences in rater consistency

Rater leniency or severity is the most commonly investigated rater error (Wolfe, 2004) and occurs when raters are consistent in how they rank performances, but do not agree in their ratings because they are either harsher or more lenient than other raters. As will be discussed in the review on rater cognition research into various rater attributes (Section 2.5.1), a general tendency towards being lenient or severe can be investigated on the individual and group level. According to Myford and Wolf (2003), there are several potential explanations for leniency/severity including: variances in individual disposition, avoidance of unpleasant judgement of others, or preferring to err in favour of a candidate. According to Cronbach (1990) leniency/severity effects must be considered as the most serious kind of error if it remains undetected because it has the potential to affect each rating decision.

Bias effects can be defined as ‘systematic subpatterns’ when raters interact with certain aspects of performances (Wigglesworth, 1993, p. 309). Raters may systematically react more harshly or severely to certain language features, traits as defined in the rating scales, or tasks (McNamara, 1996). Bias as defined by Myford and Wolfe (2003) describes score interactions between particular raters or rater groups (according to language background, beliefs, attitudes, etc.) and particular examinee groups (according to examinee gender, age, perceived language background, perceived cultural backgrounds, etc.).

One common conceptual definition of the *halo* effect is that raters “fail to distinguish between conceptually distinct features of examinee performance, but rather provide highly similar ratings across those features” (Eckes, 2009, p. 5). Thus, halo is regarded as one form of rater error. Two further definitions reflect the nature of the construct

that is being measured and the nature of the operationalizations used (Murphy et al., 1993; Myford & Wolfe, 2003). *True* (or valid) halo is to be expected if the dimensions or criteria guiding the raters overlap and are not entirely distinct. *Illusory* (or invalid) halo may be linked to various phenomena related to human perception, memory, but also rating scale design or faulty preconceptions about the connectedness of certain dimensions and leads raters to carry over ratings on one criterion to that of other criteria. While these conceptually different definitions pose a psychometric challenge, halo effects may in fact contribute positively to rater accuracy (Myford & Wolfe, 2003). Whenever raters rate via an analytic rating scale and have to provide scores for several criteria at the same time, halo effects may impact test scores and inflate correlations between scores on different dimensions (e.g., Sawaki, 2007).

Raters may also differ in how they use the rating scale. For instance, scale steps or increments may vary between raters, in that some raters may tend to make it more difficult or easier to reach a certain band on the rating scale (McNamara, 1996). Rating patterns can also reveal *central tendency* effects (i.e., tending to avoid the upper or lower ends of the scale), *extreme tendency* effects (i.e., tending to prefer the upper or lower ends of the scale), as well as *restricted range* effects (i.e., a tending to overuse certain bands of the rating scale) (Myford & Wolfe, 2003; Wolfe, 2004). These effects might be caused by rater perception when raters struggle to identify differences between the performances or tend to overemphasize the differences between them. They may also be due to individual response styles, personal preferences or cultural norms (Myford & Wolfe, 2003; Wolfe, 2004).

Finally, raters may vary in their consistency. As observed by McNamara (1996) and numerous other studies (e.g., Eckes, 2015; Lim, 2011), some raters may be more

consistent and less prone to error than others. Under particular circumstances, raters may appear erratic due to undetected effects (e.g., the quality of a particular set of performances or conditions of rating). Persistent erratic rater behaviour, however, is unpredictable and necessitates retraining or even retiring raters in standardized assessment contexts (McNamara, 1996).

Various approaches have been suggested to deal with rater variability. Before the actual test, rater training and accreditation may be *a priori* measures used to improve rating quality. Once a test is operational, using at least two raters per candidate and establishing rater reliability are common practice (Eckes, 2009; McNamara, 1996). A more sophisticated approach is to mitigate rater error through MFRM analysis and correct scores before publication of results.

2.5 Rater Cognition

In an effort to learn more about the effect of using raters in assessments and predict or explain rater variability, there has been a considerable increase of research into *rater cognition* (Bachman, 2000; Bejar, 2012; Eckes, 2015; Suto, 2012). Two main focus points of rater cognition research are (1) the effects of certain rater attributes on scores and (2) the nature of the mental processes that occur during rating (Bejar, 2012; Suto, 2012). Research into rater attributes often employs statistical analysis to identify and investigate rating patterns which may be linked to particular rater background variables (Wolfe & McVay, 2012). Studies about raters' mental processes tend to require exploratory research designs to investigate the thought processes of raters in smaller samples via qualitative methods. A selection of the main findings of these two

branches that pertain to the aims of this dissertation will be summarized in the following two sections, 2.5.1 and 2.5.2.

2.5.1 Rater Attributes

In this dissertation, the term *rater attributes* refers to particular traits raters already bring to the rating task. The term *rater characteristics* is sometimes used synonymously with rater attributes (e.g., Kang, 2012), but has also been used to describe particular scoring patterns (Lumley & McNamara, 1995) associated with groups of raters (similar to *rater types* found in Eckes, 2008, 2012).

There is a considerable number of studies investigating how rater attributes might contribute to score variability. Among the most frequently researched factors are the raters' first or second language background (e.g., Zhang & Elder, 2011, 2013), accent familiarity (e.g., Winke et al., 2013), educational background (e.g., Wiseman, 2012), training (e.g., Lim, 2011; Davis, 2008, 2015), or expertise (e.g., Barkaoui, 2010a, 2010b, 2011; Wolfe, Kao & Ranney, 1998; Zhang, 2016). More recent studies explored raters' perception of test takers' first language (Huhta et al., 2019) and social bias (Kang et al., 2019).

2.5.1.1 Language Background

The role that language background may have on L2 language raters' performance has been widely recognized and investigated particularly in speaking examinations. As there are numerous constellations between the language backgrounds of raters and test takers, studies vary considerably in how they operationalize background variables.

One branch of studies which produced mixed results investigated the rating patterns of native speaker (NS) raters compared to non-native speaker (NNS) raters (e.g.,

Brown, 1995; Kang et al., 2019; Kim, 2009; Xi & Mollaun, 2009; Zhang & Elder, 2011, 2013). When rating with a holistic rating scale, Zhang and Elder (2011) could find no systematic differences between NS and NNS raters and concluded that “the native/non-native dichotomy is not meaningful in that raters . . . rank candidates the same way” (p. 45). Their findings are similar to Kim (2009) who identified similar rating patterns in both groups, but differences in detail and elaboration as NS raters were able to provide more detail concerning the criteria pronunciation, grammar and accuracy. Wei and Llosa (2015) also detected that NS Indian raters understood Indian test takers better than the American raters, but could not establish any significant difference in use of criteria, consistency or severity. The contribution of native speaker status was also unclear in Brown (1995). In a comparatively large study of naïve raters prior to training ($N = 82$), however, Kang et al. (2019) found native speaker status to be a strong predictor, with NS being less severe in their ratings than NNS. They recommend considering native speaker status when training raters and building rater pools to reach improved inter-rater reliability.

Another variable related to language background that is explored in rater cognition research is accent familiarity. In a study including 99 IELTS examiners in five geographically dispersed test centres, Carey et al. (2011) found that prolonged exposure to accent and familiarity with test takers’ L1 led to more lenient ratings (Carey et al., 2011). This finding appears to be supported by Kang et al. (2019), who operationalized accent familiarity as time spent with NNS of English. Similarly, Park (2020) established that heritage teachers who were the most familiar with the speaker’s L1 were more lenient when assessing global proficiency and accentedness than those who were less familiar or not familiar with the test takers’ L1. As Winke et al. (2013)

have found, this leniency effect of accent familiarity can even be observed when raters only have some experience¹ of learning the test takers' L1.

2.5.1.2 Experience

There are a handful of studies that investigated the effects of rater *experience* – often distinguishing *new raters* or *novices* from *experts* – and rater *expertise* – comparing raters with varying rating competence levels. As Lim (2011) observes, rater experience and rater expertise are likely related as it is often hoped that the former might improve the latter; however, such a causal link is not necessarily a given and warrants scrutiny. Hence, the distinction will also be upheld for this dissertation. The main question driving research into the effect of rater experience is to determine the extent to which experience coincides with improved rating quality.

Writing. Barkaoui's series of articles on the effect of rater experience and rating scale type (2010a, 2010b and 2011²) is the most rigorous effort to date to investigate these two factors in the context of English essay writing assessment. Barkaoui controlled for ESL teaching experience, rating experience, postgraduate study as well as training in assessment when forming the two groups of participants. One article (2010b) reports on the commonalities and differences of the language features raters notice when rating performances analytically and holistically. While some differences in rating patterns could partly be explained by experience, others could not. When

¹ The raters' learning experience with the test takers' L1 actually ranged from less than 2 years of learning the language at school to heritage speakers.

² All three articles seem based on one larger research project. 2010a reports on a subset of participants and essays (11 novice and 14 experienced raters, 12 essays), whereas 2010b and 2011 report on different aspects of the complete data set (31 novice and 29 experienced raters, 24 essays per rater out of a larger corpus of 180 essays).

scoring holistically, novice raters more frequently referred to “discernible (and reportable)” (p. 49) linguistic characteristics of a performance but weighed the criterion argumentation related to content and ideas heavier than experienced raters who in turn focused more strongly on language and also referred to language quality on a more global level. Experienced raters were also more severe than novices when rating holistically. Other patterns of between-rater variance which were explored in the multi-level regression model suggest that rater factors which were not controlled for – Barkaoui suggests L1 or writing experience – could be the source of these differences (p. 42). When looking more closely at the decision-making behaviours and strategies, Barkaoui (2010a) found no significant differences between the experienced and novice group. An analysis of the rating behaviours by rating scale and rater group then indicates that the rating method (i.e., the rating scale type) rather than the rater’s experience shaped the features of rater cognition under investigation.

Cumming (1990), also focusing on writing, provided one of the first investigations of rater cognition and contrasted the rating behaviours of novice and experienced raters to establish typical decision-making behaviours the two groups employed. In this study, the novices ($n = 7$) and experienced raters ($n = 6$) rated written performances holistically. Raters differed in severity regarding content and rhetorical organization in that novices were more lenient, but the two groups did not differ in terms of their severity regarding language use. Cumming (1990) also found experienced raters to be more self-reflexive and employ a more encompassing combination of knowledge, self-control strategies and diverse criteria to come to a rating decision (p. 43). Novices, on the other hand tended to pay attention to a smaller range of features and developed a tendency to correct or edit the language rather than use language features as information for a grade. Similar to Barkaoui’s studies (2010a, 2010b, 2011), Cumming

(1990) found evidence of great inter-individual variation, with some novices displaying similar features as experienced raters.

Speaking. There do not appear to be many studies into the rater cognition of novice raters when assessing speaking. The most rigorous study to date may be Kim (2009, 2015), who recruited nine raters and grouped them as either novice, developing or expert ($n = 3$ each) based on their rating experience, teaching experience, rater training, and educational background. Using an analytic rating scale, the participants rated 18 ESL speaking performances across three rating sessions and their verbal report data was analysed to investigate their intra-individual differences as well as progress. Kim (2015) found differences in how raters handled the rating task, with novices initially finding it more difficult to deal with the cognitive demands and focusing on a limited range of features. The developing raters, who had previous experience, struggled to adjust to the new rating scale and showed weaker rating quality measures in the beginning. Both groups, novices and developing raters, however, managed to improve over the course of the rating sessions even without being provided feedback. Depending on their previous experience, Kim (2009, 2015) concludes, raters may have different training needs and future studies should control effectively for background variables such as rating experience or teaching experience. The study is based on a small sample of raters ($N = 9$) and relatively few performances. However, many of Kim's observations appear to match findings from Cumming (1990) and Barkaoui (2010a, 2010b, 2011).

Training. One route to achieve higher experience in raters is to train them. However, a consistent finding in many studies in both contexts, speaking and writing, is that training appears to affect some parameters of rater behaviour more than others (e.g.,

Davis, 2008, 2015; Weigle, 1998; Kang et al., 2019). As Weigle (1998) showed, training has the potential to shape scoring patterns in both, experienced and less experienced raters. It reduces random error on part of the individual rater (McNamara, 1996), leads to greater agreement between raters and accuracy in light of reference scores (Davis, 2008, 2015). Interestingly, training does not seem to exert the same impact on all features of rating behaviour. Studies including raters with previous teaching or rating experience consistently confirmed that training affects consistency more than severity (Lumley & McNamara, 1995; Weigle, 1998) and does not eradicate differences in severity (McNamara, 1996; Davis, 2008, 2015; Kim, 2011).

Overall, research into the rating behaviour of new or novice raters appears to have produced mixed results. In some studies, participants already achieve satisfactory levels of severity or consistency even before receiving training (Davis, 2008, 2015; Lim, 2011), or reach inconspicuous patterns quickly once they become operational raters and are exposed to many performances (Lim, 2011). In a comparison between new raters and experienced raters, Lim (2011) found that new raters may or may not be distinguishable from experienced raters in terms of severity or consistency. On the other hand, Weigle (1998) observed that the newer raters in her study were clearly more severe and inconsistent prior to training than the more experienced raters. However, the experienced raters in Kim's study (2011), were more severe than novice raters, but all raters were internally consistent when using an analytic rating scale even prior to training.

One possible explanation for the somewhat mixed findings regarding the effect of rater experience could lie in the operationalisation of experience and confounding background variables. In some of the studies just discussed, rater experience is

commonly defined as experience in rating with a certain scale or within a certain context, but there is often considerable variation among the participants along other variables. Kim's (2015) study of nine raters, which was particularly rigorous with respect to sampling, controlled for rating experience, teaching experience, rater training and educational background. On the other hand, the only attribute we learn about Cumming's (1990) novice raters ($N = 7$) is that they were recruited from a teaching English as a second language university class. Davies' sample ($N = 20$) consisted of experienced teachers. Weigle (1998) reported that the teaching experience of the 'new' rater group in her study ranged from 0 to 10 years, but the description of the sample is opaque about other background variables. The eleven participants in Lim (2011) were employed in a language testing context and had a linguistics background at undergraduate level. Given that tutoring experience in and of itself has been found to impact listener's perception of L2 speech (Kang, 2012), classroom teaching experience is likely a significant factor that is not controlled for or considered in some of the studies included in this review.

2.5.1.3 Expertise

Only a handful of studies investigated rater expertise by either comparing features of rater cognition between rater groups of varying accuracy and consistency (Davis, 2008, 2015; Kim, 2015; Wolfe et al., 1998; Zhang, 2016), or focusing entirely on highly proficient raters (Lumley, 2005). Predictions about expert rater behaviour are generally more coherent than predictions about experience effects (see previous section). One explanation could be that expertise, as opposed to experience, can be defined much more clearly based on actual measures of rater quality and rater performance.

Accurate raters display a range of behaviours and strategies that are different from inaccurate raters. In several studies, accurate raters were found to be more consistent in identifying and diagnosing errors (Cumming, 1990; Zhang, 2016). They exert more self-control, use self-monitoring strategies, and appear highly suspicious of bias when detecting particularly striking features, or what Zhang (2016, p. 49) called “shining points” (Cumming, 1990; Cumming et al., 2001, 2002; Wolfe, 1997; Zhang, 2016). This makes them more sensitive to uneven profiles of candidates (Zhang, 2016). Wolfe et al. (1998) argue that more proficient raters are more accurate because their judgments are closely aligned with the scale descriptors while less accurate raters relied more on self-generated descriptions. This finding was also supported by Zhang (2016) who identified idiosyncratic approaches to defining errors.

Expert raters are also able to form a fuller and more abstract representation of the performances as they emphasize integrating various aspects of language use rather than local phenomena, summarize details into more abstract categories and form more abstract overall judgements (Cumming, 1990; Wolfe, 1997; Zhang, 2016). Wolfe et al. (1998) describe this as raters being more holistic in their approach and holding back on decisions until they have read the entire text. Less accurate raters, on the other hand, were prone to an iterative bottom-up process of reading and reacting.

While inaccurate raters appear less systematic their approach of dealing with certain decisions, accurate raters adopt distinct strategies for distinguishing between bands. According to Zhang (2016), accurate raters shifted from looking at language accuracy in weaker performances to focussing on the variety and range of language in the strong performances (Zhang, 2016). Thus, they were flexible in that they employed different approaches, but also systematic into when each mode was activated.

As was presented thus far, there is a considerable number of studies into the effects of rater attributes on the assessment of writing and speaking. The effects of certain rater attributes like language background, experience and expertise converge in many studies. Predicting effects, however, remains difficult due the limited amount of rater background information many studies provide or fuzzy distinctions between expertise and experience.

2.5.2 The Rating Process

The second branch of rater cognition research pertains to studies into the rating process, which may focus on identifying the features that raters notice as they rate or attempting to map the decision-making. This section begins by reviewing rater cognition studies in the context of second language writing for two reasons. First, there are in general fewer studies on the rating of speaking – this has been pointed out before (e.g., Brown, 2005), yet little seems to have changed about this imbalance. Second, there is no model specific to the rating of speaking which conceptualizes the cognitive processes involved (Purpura, 2013). The model of speaking examinations as suggested by Fulcher (2003, see Section 2.2) is a useful framework for providing an overview of the components involved in the assessment of speaking, but as stated earlier includes only little detail concerning the rater and the rating process.

2.5.2.1 Writing

In terms of cognitive processes during the rating of written performances, several models have been developed to describe thinking patterns and the broad stages involved in rating (Cumming, 1990, Cumming et al., 2002; Milanovic et al., 1996; Lumley, 2002, 2005; Wolfe, 1997).

Cumming (1990) was the first systematic investigation of the rating process when assessing written performance. Through concurrent think-aloud protocols, Cumming mapped the decision-making of six experienced and seven novice ESL teachers. The participants rated 12 performances, but instead of using a rating scale, they were asked to rate each performance with a score from 1 to 4 along three criteria: language use, rhetorical organization, and substantive content. Cumming was able to identify 28 decision-making behaviours which were categorized as either interpretation strategies (e.g., scanning the text, interpreting ambiguous phrases, etc.) or judgement strategies (e.g., assessing the development of topics, counting propositions, etc.). Strategies were also found to either focus on the performance (i.e., content, language, organisation) or self-regulation of rater behaviour. As discussed previously in the section on rater attributes (see Section 2.5.1), Cumming (1990) found that rating behaviour was influenced by experience in that experts were able to integrate many sources of information and employed more metacognitive strategies to regulate their thinking process. Novice raters, in contrast, made use of fewer criteria and tended to depend on their general experience as readers or previous experience in related fields, for instance editing, to form their rating decisions. Cumming et al. (2002) later refined Cumming's (1990) framework and used it to research native speaker and non-native speaker rating behaviours. They suggest a prototypical sequence of decision making while rating consisting of 1) scanning the compositions for surface-level quality indicators, 2) engaging in interpretation strategies and certain judgement strategies, and 3) articulating a score while summarizing and reinterpreting observations and judgements about the performance.

Shortly after Cumming et al. (2002), Lumley (2002, 2005) provided an in-depth investigation of the rating process for writing from the perspective of four highly

experienced and proficient raters in the context of an English for specific purposes test. His study takes a slightly different approach in that he focused on how raters use and interpret the criteria and descriptors of an analytic rating scale. Lumley (2005) found that due to the exemplary and abstract nature of the rating scale, the rating process was not a straight-forward allocation of bands to a performance. Instead, the think-aloud protocols revealed that raters did not fully internalise the scale, but instead actively engaged with the descriptors for each rating. Raters were also found to employ a range of strategies to resolve the tensions between the idealised performance features as described in the rating scale and the complexities of the actual performance. As can be seen in Figure 2.3, Lumley (2005) proposes a highly detailed model of the rating process which consists of three stages (i.e., reading, scoring, and conclusion) but also acknowledges that the rating process is closely linked to the goals and requirements of the institution (i.e., the institutional level) and the capacities and needs of the individual (i.e., the interpretational level). These observations led Lumley (2005) to suggest that the rating process never becomes fully automated but remains complex and challenging even for highly experienced raters such as the ones included in his study.

Figure 2.3 Model of the rating process by Lumley (2005)



A third perspective on the rating process is Wolfe's cognitive model (1997, 2006; Wolfe et al., 1998). This model is different to Cumming (1990) and Lumley (2005) in that it leans towards concepts established in cognitive sciences. According to Wolfe's model, the rating process is shaped by two cognitive components, a *framework of scoring* and a *framework of writing*. The framework of scoring constitutes the "processing actions" raters engage in as they read and evaluate a performance. The framework is guided by the rater's concept of the rating task, i.e., their understanding of how a candidate performance is to be read, interpreted, evaluated, and used to justify rating decisions. The framework of writing constitutes the rater's mental representation of the features that are typical of "proficient and non-proficient writing" (p. 90). The true quality of the framework of writing cannot be observed directly but must be inferred from the scores. Aspects such as rater training, experience and the rating scale in use may shape the framework of writing and according to Wolf, the

frameworks of writing of reliable raters should be expected to be in close alignment with the rating scale.

The last model to be included here is Bejar's descriptive model (2012) which appears to integrate many of the features from Cumming (1990), Wolfe (1997), Cumming et al. (2002) and Lumley (2005) (see Figure 2.3). Bejar's model differentiates between the assessment design phase and the scoring phase. During the assessment design phase, the assessment is conceptualized, scales (or scoring rubrics if following Bejar's terminology), tasks as well as procedures and benchmarks are created, and raters are recruited and trained to "form a mental scoring rubric" (p. 5). The scoring phase follows the design phase and consists of three broad stages: forming a representation of the performance, comparing the representation of the scale with the representation of the performance, and assigning a score. Similar to Fulcher's (2003) and Eckes' (2015) componential models, Bejar's descriptive model also specifies the numerous factors (e.g., training, rating conditions, sequence of performances, etc.) which might bear an influence on the score.

Bejar's model is somewhat more general in that it does not specifically describe the process of rating spoken nor written language. However, as the context of writing appears foregrounded and implied throughout and aspects of rating speaking are not considered it will feature as the last model in this section.

Figure 2.4 Bejar's (2012) descriptive model of the rating process

Assessment design phase	Scoring phase
<ul style="list-style-type: none"> Assessment design process identifies evidence of the different levels performance called for. Scoring rubrics are formulated to formalize the relevant gradations of performance. Items and tasks are created to elicit performance. Items and tasks are pretested and evaluated, do they elicit the evidence called for? If so, collect benchmark and rangefinders to train raters. Recruit and train raters using the scoring rubric, benchmark and range finders. Raters form a mental scoring rubric based on the training Ideally, as a result of the training all raters will rate equivalently but their background or other factors could lead to rater effects 	<ul style="list-style-type: none"> Rater reads a work product and forms a mental response representation Rater compares the similarity of resulting representation with mental scoring rubric Based on that comparison the rater tentatively assigns the response to a score category The score they assign to a specific response depends on <ul style="list-style-type: none"> The true quality of the response The quality of their mental scoring rubric, The quality of the representation they formed for this response The prior information they have accumulated during the scoring. The state of the rater, e.g., fatigue Environmental conditions The nature of the responses previously scored

2.5.2.2 Speaking

Compared to writing, there are generally fewer studies investigating the process of rating of L2 speaking (Brown et al., 2005; Purpura, 2013), which may partly be due to methodological challenges. A commonly used method to research rating processes in writing is the concurrent think-aloud method (Bejar, 2012), which would interfere with the real-time nature of assessing speaking and interrupt authentic rating processes (Brown et al., 2005). As such, many rater cognition studies in the context of speaking focus on identifying the features raters attend to rather than the mental processes during rating.

Upon investigating the features raters notice in oral proficiency interviews, Pollit and Murray (1996) found that the six raters adapted their criteria depending on their perceived ability of the speakers. In this study, raters were not provided with a rating scale but had to rate the performances in sets of two, following Thurstone's method of paired comparisons. The results showed that raters selected salient features based on the weaker speaker in each pair. Whenever the speakers' proficiency was higher,

raters shifted their attention from features of grammatical competence (defined by the authors as grammar, vocabulary, comprehension, and pronunciation) to features of discourse competence (defined as stylistic devices).

Investigating the features raters attend to can also be used to inform rating scale design or validate new modes of assessment. Brown et al. (2005) used comments generated by raters to help create descriptions of speaking proficiency at different levels. This study also showed that raters adapted their focus depending on task type (independent or integrated tasks) and potentially even to different tasks. Nakatsuhara et al. (2020) compared rating behaviour under live, audio and video condition and analysed written justifications and verbal reports from six certified IELTS raters. They found that raters noticed more positive features when rating live performances and were more severe when rating audio recordings. This might be due to the rich information raters take in about a performance that is not conveyed in the audio recording, and both aids their comprehension and interpretation of test-taker language.

Several studies have shown that the features raters notice may vary considerably even though rating scales are provided. Meiron (1998, as cited in Brown et al., 2005) and Brown (2000) found evidence that raters may resort to self-generated features not included in the rating scales. As raters try to come to a judgement about a candidate's ability and in the face of the limited information that can be gathered through observing the performance, raters may resort to making inferences about the speakers' ability or personality (Brown, 2000). According to Brown (2000), raters also varied in their approach to assessing task completion. While some raters employed a narrow approach and looked for specific features (i.e., certain grammatical cues), others

emphasized the general appropriateness of a speaker's performance to the context of the task.

Two influential studies (Orr, 2002; May, 2006) investigated which features raters attend to when rating paired speaking tests. In Orr (2002), 32 certified raters rated two paired First Certificate in English performances (i.e., four candidates) and subsequently produced a verbal report justifying the scores they awarded. The results showed that raters awarding the same rating may differ considerably in how they perceived a performance. Orr (2002) argues that this may be due to varying rater severity, but also incorrect scale use. Another finding was that raters generally struggled to confine their comments to features identified in the rating scale and that construct-irrelevant observations varied between raters. Orr (2002) presents a particularly overt example of how much rater perception may diverge even when raters have received training and are provided with a rating scale. May (2006) specifically targeted the construct of interactional competence as raters appear to comprehend and apply it to rating performance in paired discussion tasks. Combining the analysis of rater notes, summary statements, verbal reports and rater discussions, May was able to show the degree to which raters had to "flesh out" the criteria they were provided to come to rating decisions. Similar to other studies (e.g., Brown, 2000; Orr, 2002), raters were found to notice non-linguistic features and integrate these with the descriptors provided in the scale.

While the previous studies investigated rater cognition in paired or interactive tasks, Davis (2008, 2015) exclusively focussed on rater cognition when rating individual long turns. Davis tallied and compared the comments from nine raters who were grouped according to several rating measures (accuracy, severity, fit) into a proficient,

non-proficient and developing group. When comparing the comments, Davis (2015) found no significant differences between the groups and no evidence for change over the course of the rating sessions. However, the small group sizes of three individuals per group and the broad operationalisation may have failed to capture differences between the groups.

The findings presented thus far relate to the features raters notice when assessing spoken language. There are also some findings concerning the mental processes that may be activated in the context of speaking assessment. Pollit and Murray (1996) found in their study of paired comparison rating that raters employed two broad approaches, which appears like what Meiron (1998, as cited in Brown et al., 2005) seems to have discovered. Three of the six raters tended to follow an intuitive or *synthetic* approach, in that they would form an overall first impression of the speaker based on a “primary indicator” (p. 5) of that level. In their memory of the performance, they would tend to conglomerate certain traits which they felt representative of the test taker level. The other half of the rater group tended to follow what the authors labelled a less “natural” approach (p. 5). Even though the paired comparison method favoured a more holistic approach to assessing performances, these raters would tally observations about the candidates as the performance unfolded and then come to a judgement which they reconciled with the rating method.

More recently, Davis (2008, 2015) made a significant contribution by focussing on the effects of rater training and the rating process in general. In his study, 20 experienced English teachers were trained with a holistic rating scale. In a remote data collection setup, each participant rated 100 performances prior to the first training session, followed by three more rating sessions of 100 performances each. Davis

(2008, 2015) also collected time stamps which recorded the moment each rating was submitted through embedding JavaScript in Adobe rating forms. Over the course of the rating sessions, raters improved most significantly after the first training sessions. After that, raters continued to improve as far as their accuracy was concerned, but their inter-rater reliability remained relatively unaffected by their increase in experience. In addition, Davis (2008, 2015) observed that more proficient raters would require more time for taking their rating decisions and made fewer miscellaneous or unrelated comments about the performances and fewer comments about the scoring process as such. However, as the data collection was carried out remotely, Davis could not explain whether the accurate raters needed longer for their decisions because of certain strategies, like listening to a performance twice or using exemplars, and the set-up of the study also did not reveal which strategies or meta-cognitive strategies the raters might have employed.

2.5.3 The Role of Rating Scales

As identified earlier in this chapter, rating scales are a crucial element of performance-based language assessment and considered a fundamental prerequisite of reliable and valid assessment. In the Cambridge Dictionary of Language Testing scales are defined as: “consisting of a series of constructed levels against which a language learner’s performance is judged” (Davies et al., 1999). Rating scales serve multiple purposes; they are operationalizations of the construct (Fulcher, 2003), and as such reflect what testing institutions value in performances (Knoch, 2009; Lumley, 2005), and provide transparency towards language learners and other stakeholders (Davis, 2018). Rating scales consist of several descriptors which describe expected performance features at different ability levels. These descriptors are “the most specific operationalisation” of

the construct to be assessed (Harsch & Martin, 2013, p. 287), and define the inferences that can be made about a speaker's performance (Fulcher, 2003).

Rating scales may be characterized along several traits. Alderson (1991) established how rating scale design may vary depending on the intended users. He distinguished between *user-oriented*, *constructor-oriented*, and *assessor-oriented* scales. User-oriented scales provide information about typical expected behaviours at specific levels and may be phrased in terms of what candidates can do with the language. Constructor-oriented scales may be very detailed and provide important additional information concerning the tasks. Assessor-oriented scales include salient information for the examiners or raters in a more compressed format so that they can be accessed and processed easily during the exam. Thus, as Alderson (1991) claims, depending on the target audience and purpose, rating scales may be designed with different levels of detail.

The two main approaches to rating language performances have led to the development of two distinct types: *holistic* rating and *analytic* (e.g., Harsch and Martin, 2013). A third approach to scale design, which will not be discussed at length in this review is to select and focus on a task-dependent feature. *Primary-trait scales* are created by identifying a key language feature and subsequent rating decisions are based on observations regarding this one feature in the performance (Davis, 2018).

Each type of rating scale comes with its own set of 'implications' for validity and reliability of scores (Harsch & Martin, 2013) and may, thus, be suitable for different test purposes, assessment contexts and raters (Barkaoui, 2011). As holistic rating requires only "a single overall judgement" (Davis, 2018, p. 1), it is a faster and more economical method (Weigle, 2002). A shortcoming of holistic rating is that the score

is difficult to interpret as raters may be found to award similar scores for different reasons (Harsch & Martin, 2012; Weigle, 2002). Furthermore, providing only a single score veils the complexity of the construct (Fulcher, 2003) which in turn is a threat for scoring validity (Weir, 2005). Also, one feature might easily dominate the overall impression leading to undetected halo effects. As Huot (1990) points out, an emphasis on obtaining high inter-rater agreement indices may have contributed to the widespread use of holistic marking at the expense of validity. The current view on holistic rating appears to be that it may be economical but associated with lower inter-rater reliability than the analytic method (Weigle, 2002). According to Weigle (2002, p. 73), Weir (1990) also identified several studies which found that holistic rating scales produced less reliable scores than analytic scales.

Analytic scales, on the other hand are sensitive to various constructs and have the potential to offer diagnostic feedback (Fulcher, 2003), which makes them more relevant in pedagogical contexts (Hughes, 2003). Particularly with novices, analytic scales may help raters focus on a broad range of features (Barkaoui, 2011; Harsch & Martin, 2013) as they draw “raters’ attention to specific aspects of students’ performances” (Cumming, 1990, p. 42). However, users may struggle to differentiate between the constructs operationalized in the criteria which may incidentally also create halo effects (Cumming, 1990; Fulcher, 2003). Interestingly, one aspect that is hardly discussed in the literature is the extent to which the mode of delivery of the test performance (i.e., assessing written language vs. assessing spoken language) might alter the cognitive requirements of the rating task. As numerous rater cognition studies found (e.g., Cumming et al., 2002), raters often reread the performances or revisit certain features when they take their rating decisions. A spoken performance, on the other hand, cannot be navigated in this way and raters may have to employ certain

strategies to mitigate the effect this may have on their capacity to process both, the performance and the rating scale.

Regardless of which scale is being used, raters have been shown to struggle with relating performances to rating scale descriptors for a whole range of reasons. One challenging aspect may be that performances can be uneven as candidates may be nervous or overly confident in the beginning and shift throughout the performance (Underhill, 1987), or that a descriptor does not adequately capture the nature of the performance (Alderson, 1991). A frequent criticism towards rating scales is that they may also lack empirical footing (Weir, 2005; Fulcher, 2003), and even if scales are created in an empirical process, they may suggest a stepwise progression that does not represent natural acquisition patterns (Fulcher, 2003). Finally, striking a balance between a detailed enough description and usability remains a challenge; if the scales provide too little detail, raters struggle to locate a performance on the spectrum described by the scale, while adding more detail to the scales may render them unwieldy (Alderson, 1991; Underhill, 1987).

Barkaoui (2010b), investigated the effect of rating scale type on novice and expert raters by comparing their think aloud protocols when rating written performances analytically and holistically. His findings showed that the scale type had a larger effect on rater cognition than the level of experience. Holistic rating led to more engagement in interpretation strategies, which focus on the performance, and analytic rating induced more judgement strategies and engagement with the scale as several separate decisions need to be formed for each performance. In another quantitative study involving a considerably larger sample, Barkaoui (2011) found that raters were more lenient, more consistent and distinguished more clearly between the different ability

levels when they rated analytically. More recently, eye-tracking studies have found evidence that not just the rating scale type, but also the format might bear an influence on rater cognition (Winke & Lim, 2015; Ballard, 2017). According to these studies, rating scale layout may lead to raters paying more attention to features that are presented towards the left of the scale than to the right (Ballard, 2017; Winke & Lim, 2015).

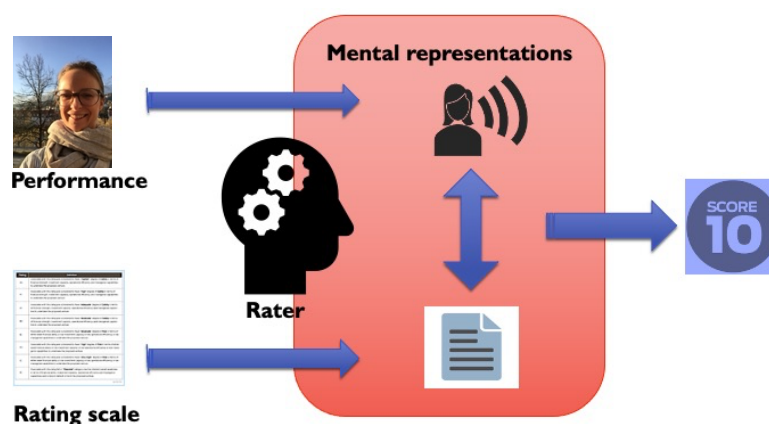
A final concern regarding rating scale design and use which needs to be considered in connection with rater cognition is whether the scale is oriented towards a specific task or not. Primary-trait scales are *task-dependent* and need to be developed for each new task (Davis, 2018). Fulcher (2017), however, observes that the more general and *task-independent* rating scales are in their wording (see also distinction of *producer-oriented* and *assessor-oriented*), the more difficult it may become to apply the descriptors to actual language performances. On the upside, task-independent rating scales allow score users to make more general inferences about the speaker's ability and may be more appropriate whenever general language proficiency needs to be assessed (Davis, 2018).

2.5.4 Summary Rater Cognition

This review of rater cognition studies has shown that the field has developed and sustained an interest into the process of rating and how these may be shaped by specific rater attributes as well as attributes of the task or performance. Rater attributes such as language background and experience were identified as important factors and investigated in various configurations, and the rating process was observed in the context of writing as well as speaking assessments.

Figure 2.5 is meant to summarise those features that several models (Bejar, 2012; McNamara, 1996; and Wolfe et al., 1998) appear to share about the rating process. The rater and their attributes are at the heart of these models. During a speaking examination, raters continuously form a representation of the speaking performance as it unfolds and compare it to their mental representation of the rating scale. The performance influences the features of the rating scale that are recalled or looked for in the scale, and the rating scale influences the features that raters look for or attend to in a performance. The quality of the representations that raters can form of both, performance and rating scale, depends on context-independent factors (e.g., the quality of their training, their experience, languages spoken, accent familiarity), as well as context-dependent factors (e.g., conditions of rating, fatigue, stakes of examination).

Figure 2.5 Simplified model of rating process (based on Bejar, 2012; McNamara, 1996; Wolfe et al., 1998)



Despite the growing interest in rater cognition, the studies and findings presented here are limited in certain respects. For instance, it is remarkable that there is a host of studies into rater attributes, but fewer studies that attempt to link rater attributes to observable rating behaviours, particularly in the context of speaking. As Kim (2015)

argues, “raters’ rating patterns have been analysed excessively without efforts to understand who they are and what they bring to the assessment contexts” (p. 251).

Another limiting feature appears to be the scope of settings in which rater cognition research is carried out. Often, studies focus on highly standardized tests (e.g., Davis, 2008; Eckes, 2012; Lim, 2008; Winke et al., 2013) and research efforts are driven by a need to validate test instruments or procedures directly related to rating or training (e.g., Eckes, 2012; McNamara, 1996). While fulfilling an important purpose for test developers and stakeholders, this research only partly addresses broader issues related to rater cognition which may be relevant in various rating contexts.

Finally, as Kang (2012) points out, there is a range of rater attributes (e.g., attitudes or beliefs) which have largely been ignored but have the potential to explain score variance in language tests as well as some of the mixed findings within rater cognition research. Given the nature of rating spoken language, which involves extensive real-time and multi-modal speech processing, it seems theoretically unfounded to generalize from research into the rating of writing to the rating of speaking. When assessing written language, the raters can read and reread the text while deliberating their decisions, but when rating spoken language, the rater is under a constant pressure to form a full representation of the speaker’s performance and compare it to their understanding of the criteria in the rating scale. This is even more the case when performances are not recorded and can only be experienced in the moment of the assessment, or when the rater also acts as interlocutor (Taylor & Galaczi, 2011). Whereas the relationship between cognitive processes and candidate performance is increasingly investigated in language testing research (e.g., Brunfaut & Révész, 2015; De Jong et al., 2012; Indrarathne & Kormos, 2017), there have been only few attempts

to explore *rater* cognition from a cognitive processing point of view (Purpura, 2013). The complexity of the rating process in the context of speaking raises the question of how raters cope with the cognitive load of real-time decision making. Insights and tools from cognitive psychology as well as judgement and decision-making research which will be reviewed in the following section as they may contribute to a refined conceptualization of the rating process while also broadening the repertoire of operationalizations and methodology.

2.6 Rater Cognition from the Perspective of Information Processing and Decision Making

Judgement and decision-making (henceforth JDM) research is concerned with learning more about how humans process information and take decisions. Thus, JDM is interdisciplinary and builds on various theories, mainly from the field of psychology and economics to model and explain decision-making processes for specific tasks or describe general decision-making patterns. Recently other fields, like health sciences, which are also concerned with rater variability have started to investigate rating from a JDM perspective. These studies have explored the schemas raters might employ or how cognitive load shapes decision processes (see St-Onge et al., 2016). It is beyond the scope of this dissertation to contextualize each of the concepts that are relevant to human decision-making processes. The following section will therefore briefly describe some key concepts of JDM relevant to rater cognition.

2.6.1 Key Concepts in JDM

Human information processing. Information processing is a cognitive approach to studying human thinking and behaviour. Wickens (1992) created one of the most widely used models of information processing and incorporated components and processes that are commonly agreed on by cognitive psychologists. Events or information in our environment are identified through human senses and stored. Next, a selection of the raw data that has been registered is then automatically and rapidly processed – in situations requiring instinctive reactions a response is instantly selected. In most other situations, perceived information is processed to be stored in long term memory or “thought about” in reasoning processes that are also conscious and accessible to the individual. Reasoning in turn requires effort and attention and cognitive functions can be disrupted or affected by emotions or stress. Responses are selected and executed through the motor system once a situation has been perceived and assessed. Feedback is constantly received for any action taken and can be intrinsic as well as extrinsic. Attention is necessary whenever processes are not fully automated and attentional resources regulate which processes are allocated the limited resources.

Figure 2.6 Wickens' model of information processing



Automated versus deliberate thinking. When humans process information and engage in complex problems, they rely on automatically processed perceptions and deliberate thought. Thought processes can be classified as quick, parallel, automated, and subconscious on the one hand (*system 1*), and deliberate, rational, effortful, and based on reasoning on the other hand (*system 2*). While humans are capable of handling numerous parallel system 1 processes (Kahneman & Frederick, 2005; Glöckner & Betsch, 2008), higher-level cognitive activities from system 2 place a strong demand on our attention (Kahneman & Frederick, 2005). Dual-processing theory tries to establish how decision-making processes are shaped by this complex interplay of system 1 and 2 (Kahneman & Frederick, 2005; Newell & Bröder, 2008), and investigates potential sources of error. Among the issues investigated is how our final choices might be impacted by our initial intuition (Simon, 2004), the sequence of processing (Russo & Doshier, 1983), or the processing effort directed towards different options or attributes of options (Reisen, Hoffrage, & Mast, 2008).

Working memory. Working memory decisively shapes the way in which complex judgement is formed because our capacity to hold, retrieve and manipulate information while thinking about a problem lies at the core of our ability to engage in decision-making at all (Newell & Bröder, 2008). The current and most widely used model rests on the work of Baddeley (2000, 2003, 2012) and Baddeley & Hitch (1974) who replaced the previously used short term memory with the term *working memory*. The model originally proposed three core components: the central executive, and the two slave systems, namely the visual-spatial sketchbook, concerned with memorizing or visualizing visual information, and the phonological loop. Baddeley (2000) later introduced a fourth component: the episodic memory.

The central executive is often viewed as the heart of the model as it regulates the slave systems, has only limited capacity, controls access to long term memory, allocates attention, and selects what may be of interest (Schellig et al., 2009). It is this reduced capacity of the executive control which explains why conscious and verbal processes are serial and disrupt one another and other processes are automated and may run in parallel. The central executive comprises three basal or lower-level regulation processes: (1) the *shifting* of attention from one task to another which is essential for switching between tasks and a prerequisite of focusing our attention on internal or external stimuli; (2) *updating* and *monitoring* which entails being able to update, manipulate and replace information that is currently held within our working memory with new information; and (3) *inhibition* which is being able to inhibit a pre-emptive reaction (e.g., Schellig, 2009; Miyake et al., 2000). The phonological loop, a slave system to the central executive, is responsible for processing language. It is considered a predictor for L1 acquisition in children and L2 acquisition in adults (Baddeley, 2003), and essential for language production and reception (Jacquemot & Scott, 2006).

Attention. As attributed in the Wickens' model of human information processing, attention takes on a specific overarching role. Attentional control is essential whenever switching in between stimuli or tasks and might bear an influence on L2 listening comprehension (Wallace, 2020). Schellig et al. (2009) list attention as a separate component of cognition, next to memory and executive functions.

The key concepts described above, namely information processing, automated and deliberate thinking, memory, and attention likely impact on decision-making as they provide the cognitive infrastructure for complex thinking. It is therefore not surprising that some decision-making research finds itself at the crossroads between the field of

JDM and cognitive psychology (e.g., Cui et al., 2015). Studies in some contexts have shown that features of working memory act as predictors for the ability to engage in rational thinking and decision-making (e.g., Fletcher et al., 2012).

When raters engage in the rating of speaking, these key concepts may also contribute to the rating process. As Bejar (2012) and Wolfe (2006) suggest, rating language performance requires raters to compare the mental representation of the performance they form during reading or listening with what they believe are salient features according to the rating scale. First, this requires raters to be able to focus on what the speakers say and inhibit thoughts that distract them from the task of rating. Raters also need to constantly update their understanding of the content of what speakers are trying to say as well as keep track of language features which might require dividing or shifting their attention. Rating spoken language is also a multi-modal activity in that it entails that raters listen to the performance, think about the performance, read the descriptors, and write down observations during or after the performance. This combination of processes poses high demands in terms of divided attention, effective task switching and meta-cognitive control.

2.6.2 Modelling Decision-making Processes

JDM research is not fully integrated in cognitive psychology. Rather, processing models such as the one proposed by Wickens form the “overarching metaphor” (Newell & Bröder, 2008, p. 196) for numerous JDM models. JDM research does not necessarily agree as to how multi-attribute decision making processes can best be modelled. In its most basic form, multi-attribute decision making can be defined as “how we make judgements when faced with multiple pieces of information” (Newell & Bröder, 2008, p. 196). However, depending on the emergence of new theories or

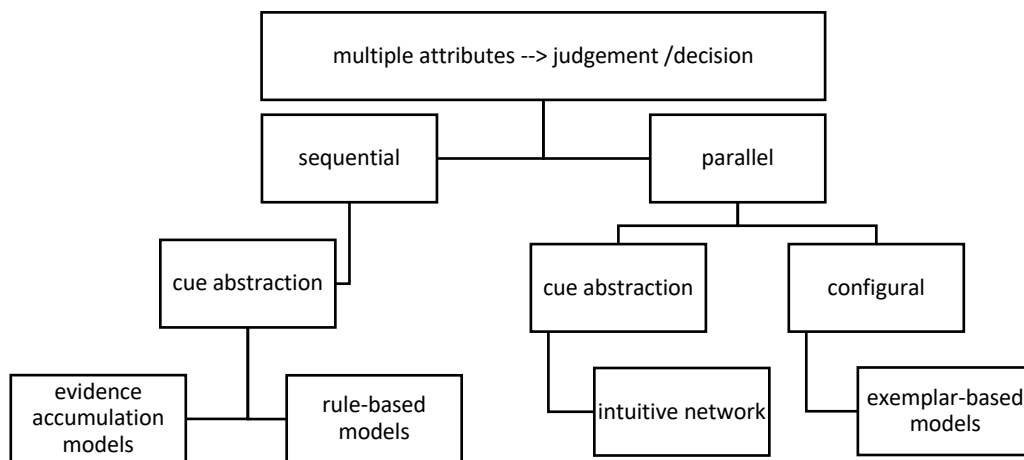
the usefulness of different theories in different contexts, the models or metaphors of decision-making processes may vary.

The key characteristics that are salient to cognitive models in general also apply to the more specific context of multi-attribute JDM. These characteristics are: (1) the mind's capacity and its limitations when processing information; (2) the nature of automatic (system 2) and controlled processes (system 1) and how they are linked to memory; (3) that humans learn from prior experiences; (4) the ability to categorise and distinguish cues to judge objects or situations; and (5) how cognition is regulated (Newell & Bröder, 2008).

In their overview article, Newell and Bröder (2008), distinguish current decision-making models on two levels: 1) how information is being processed (*sequential* or *parallel*), and 2) how cues about the object or situation to be judged are integrated into the decision process (via *cue abstraction* or *configural* cue application) (see Figure 2.7). First, information processing can be characterized as either quick, parallel, subconscious, and automated (system 2) or deliberate, rational, and effortful (system 1), and while many automated processes may run parallel without perceived effort, effortful processes “tend to disrupt each other” (Kahneman & Frederick, 2005, p. 268) and produce flawed decisions. The second level distinguishes two ways of integrating cues which are “discriminating piece(s) of information” (Läge & Hausmann, 2008, p. 229) and objects or situations can be assessed either by perceiving, weighing, and adding cues in a one-by-one fashion (*cue abstraction* processing), or considered as a non-linear network of cues in a holistic manner (*configural* processing) (Kahneman & Frederick, 2005; Newell & Bröder, 2008). Research suggests that face recognition is an example of configural processing.

Rule-based models constitute the “normative” approach to JDM and typically involve comparing decision processes and results with standards or benchmarks. This model assumes that humans are rational beings and “have orderly preferences that obey a few simple and intuitive axioms” (LeBoeuf et al., 2005, p. 243). Rational decisions express the interaction of the utility we give to certain cues and should, thus, reveal our rational preferences. Normative, rule-based models help describe sequences of actions, how people typically behave (Suto, 2012), and in what ways decisions might deviate from predictions (*biases*) to provide strategies to mitigate such bias effects (Newell & Bröder, 2008). However, such models fail to describe *how* decisions are actually made and lack predictive power (LeBoeuf et al., 2005; Glöckner & Betsch, 2008).

Figure 2.7 Multi-attribute decision models (Newell & Bröder, 2008, p. 200)



In **evidence accumulation models**, decisions are taken based on weighting perceived cues according to certain cue-criterion relations. While traditional models postulate that rational decision making is based on considering and integrating *all* pieces of information, Gigerenzer and Goldstein (1999, in Läge & Hausmann, 2008) were able to show that evidence accumulation may stop after finding just one discriminating cue.

Therefore, one cue that is perceived as valid enough might suffice to provide the confidence to take a decision, while in other situations many pieces of information are considered before information search is terminated. The issue at the heart of evidence accumulation models is to determine the threshold at which the searching for more cues stops (Läge & Hausmann, 2008) and how this threshold may vary depending on the individually desired level of confidence and other factors such as time pressure, cognitive load or stakes involved. As processing information requires effort, decision makers must feel motivated to “strike a balance between minimising their processing efforts and maximising their judgemental confidence” (Läge & Hausmann, 2008, 238).

The other two metaphors acknowledge the role of automatic and intuitive processes on a theoretical level and represent a recent interest in intuitive decision-making processes (Newell & Bröder, 2008; Zander et al., 2016).

Intuitive network models rest on the premise that humans are capable of complex processing without actively investing in conscious reasoning. As Newell and Bröder (2008, p. 201) put it: “complex process (sic.) do not necessarily imply the consumption of conscious resources or much processing time”, and, in turn, labelling heuristics to be “simple” may not adequately reflect the complexity of the processes leading to an intuitive judgement. Zander et al. (2016) were able to show on a neurological level how intuitive decision making is a gradual process in which conscious hunches emerge and are developed from preliminary and pre-conscious perceptions of patterns. Such decision-making processes are particularly prevalent in situations where choices must be made quickly or without understanding or being able to access all the information necessary.

Exemplar-based models explain why decision making may under certain conditions forgo the costly integration of numerous cues (see evidence accumulation). Instead, typical constellations of attributes stored as sets in the memory of a decision maker are activated and retrieved to take a quick decision (Newell & Bröder, 2008). While exemplar-based inference used to be considered a side effect of experience, there is empirical evidence that it is a choice or *strategy shift* (Karlsson et al., 2008) that tends to be adopted more frequently when decision-makers presented with binary options and problems that do not support additive cue-combination. When presented with rich feedback individuals would favour cue-abstraction if they were successful in unlocking the underlying cue combination rules.

The adaptive toolbox, or contingency model assumes that decision makers will choose and apply the best suited heuristic or strategy (Bröder & Newell, 2008; Newell & Bröder, 2008). Each heuristic or strategy is associated with effort or nominal cost. Thus, the adaptive toolbox model postulates that decision makers can decide on the strategy they want to apply and the effort they want to invest. As a consequence, strategies will vary depending on the information available and stakes involved: “Hence, our decisions are sometimes frugal ... and sometimes more opulent” (Newell & Bröder, 2008, p. 200). In a series of experiments, Bröder and Newell (2008) demonstrated that (1) participants can deliberately choose between different heuristics, and (2) the execution of heuristics can become routinised and less susceptible to general cognitive load. However, their study also indicates that routinisation comes at a cost of flexibility in that participants would tend to hold on to executing previously routinised heuristics even after parameters had been changed and other strategies might produce better results.

These brief summaries of the five most prevalent metaphors used to conceptualize multi-attribute decision-making processes provide a crude glimpse at the various ways in which complex decision outcomes and decision processes might differ one from the other. As Newell and Bröder (2008) emphasize, the distinctions between fast and slow processing in complex decision processes might not be absolute. Instead, decision-making processes are likely shaped by the interplay of the individual (and their capacities) and the characteristics of the decision task. Furthermore, the notion that individuals not only vary in the content but also in the process of their decision-making is gaining momentum within the JDM field and cognitive psychology (Payne & Venkatraman, 2011). While the decision-making models provide useful metaphors for the research of observable behaviours in controlled experiments and environments, a firm theoretical basis is necessary before possible behavioural patterns can be modelled and tested. From an applied and practical perspective, the propensity for certain cognitive patterns has also been researched and operationalized as an individual trait or state. Two approaches to investigating individual differences in decision-making processes will be presented in the following section.

2.6.3 Individual Differences in Decision-making Processes

2.6.3.1 Rational-experiential Inventory (REI-40)

Cognitive-experiential self-theory (CEST) rests on the premise that there are two independent modes of processing – one rational and one intuitive/experiential – and that humans can access information about their own thinking. The origins of CEST go as far back as Freud's theories about the *primary mode* (or unconscious) and the *secondary mode* (or logic, realistic) (Epstein, 1994). However, while Freud postulated that in absence of suppression all information is conscious, newer theories embrace

the notion of the *cognitive unconscious* – which are the numerous parallel processes which “automatically, effortlessly, and intuitively organize experience and direct behaviour” (Epstein, 1994, p. 710). CEST supports the dual-processing model and stipulates “that people apprehend reality in two fundamentally different ways, one variously labelled intuitive, automatic, natural, non-verbal, narrative and experiential, and the other analytical, deliberative, verbal, and rational” (Epstein, 1994, p. 710). While many people may be aware of occasional conflicts between the heart and the mind, the contributions of the two processing modes in individual decision making may vary considerably, ranging from one completely overriding the other to equal balance between the two (Pacini & Epstein, 1999).

The Rational-Experiential Inventory (REI) is one of several operationalisations of CEST (Pacini & Epstein, 1999). The original REI-59 scale includes two dimensions: the rational dimension adopted from the validated Need for Cognition (NFC) scale by Cacioppo and Petty (1982), and the experiential scale created over a period of several years and finally named Faith in Intuition scale (Epstein et al., 1996). The REI structure with two unipolar dimensions rather than one single bipolar dimension was established by Epstein et al. (1996). Studies such as, for example, Akinci and Sadler-Smith (2013) confirmed the two unipolar scales in a sample of police officers and police staff. The REI-40 (Pacini & Epstein, 1999) constitutes a further refinement of the original scale and consists of two main scales (experiential and rational processing) and two subscales within each of the main scales (ability and engagement).

The construct and structure of the REI has been confirmed in numerous validation studies. For instance, the REI-40 was mapped against and compared to the Big Five and the Preference for Intuition or Deliberation scale (Goldberg, 1990, in Wittemann

et al., 2009). Wittemann et al. (2009) confirmed the predictive validity and cross-cultural validity of the scale in the European context. In a sample of police officers and police administration staff, Akinci and Sadler-Smith (2013) found significant differences in processing preference by job type, but not by age or gender. Similarly, in a study of decision making in student pharmacists, McLaughlin et al. (2014) found no significant differences in terms of gender, race or prior degrees. Overall, the student pharmacists displayed a preference towards rational processing compared to experiential decision making, however, that difference was less pronounced in older students. In British samples including subsamples of the general population and undergraduates, Handley et al. (2000) found significant gender effects with women showing a stronger experiential preference. Their findings confirm the two unipolar scales and the subscales for rationality, but not for experientiality. Furthermore, they could not establish a link between the REI and measures of intelligence. Handley et al. (2000) conclude that the construct of the experientiality scale remains unclear.

One expectation on the basis of CEST and dual-processing is that more reflective and rational thinking leads to correct and rational decisions and intuitive processing is prone to lead to bias (Epstein, 1994). While rational thinkers enjoy engaging in rational thinking and have confidence in their decisions, others have a preference for relying on their instincts, and yet a third group might be uncomfortable about either processing mode (Fletcher et al. 2012; Pacini & Epstein, 1999). Gunnell and Ceci (2010) used the REI-40 in a study of juror cognition and hypothesized that one processing mode is likely to override the other in legal judgement contexts. They were able to show that individuals preferring the experiential mode were prone to biased judgement and heuristic thinking. Fletcher et al. (2011) provided evidence that working memory capacity directly and indirectly – via a preference for rational

processing – impacted performance on reasoning tasks. Fletcher et al. (2012) used the REI-40 to form four participant groups: *rationally dominant*, *experientially dominant*, *dual preference*, and *disengaged*. They confirmed earlier findings from Fletcher et al. (2011) and also found that participants from the *dual preference* group performed best in comparison to the other groups in a whole range of cognitive tasks (Fletcher et al., 2012). Phillips et al. (2015) confirm the role of working memory in a meta-analysis, but they also found that the predictive power of scales like the REI was task dependent in that rational processors emerge as more successful in some tasks but not in others. Similar to Fletcher et al. (2012), they conclude that being able to adapt the processing mode to the demands of the task at hand rather than an overriding preference for the rational mode might be a key to successful decision making.

2.6.3.2 General decision-making style inventory (GDMSI)

Differentiating and investigating decision-making styles is regarded as a specific field within JDM research and with a focus on professional behaviour. In response to a perceived lack of validated instruments, Scott and Bruce (1995) developed a decision-making style questionnaire (the General Decision-Making Style inventory, GDMSI). They define decision-making styles “as the learned, habitual response pattern exhibited by an individual when confronted with a decision situation. It is not a personality trait, but a habit-based propensity to react in a certain way in a specific decision context” (Scott & Bruce, 1995, p. 820). As discussed in the previous section, the theoretical backbone informing the development of the REI can be traced back to Freud’s theories. The GDMSI on the other hand, builds on Jung’s typology of attitudes and functions, and in particular on the perception functions (i.e., sensing and intuition) and judgement functions (i.e., thinking and feeling) (see Thunholm, 2004).

The GDMSI differentiates five major styles: (1) *rational* (reliant on logical evaluation of alternatives and thorough search of information), (2) *intuitive* (reliant on feeling or hunches), (3) *dependent* (seeking advice and direction from others), (4) *avoidant* (trying not to take a decision or to postpone decisions), and (5) *spontaneous* (tending to take decisions quickly). The initial iteration of the GDMSI did not include the spontaneous scale. However, as the authors explain, this style emerged as an additional factor from the first data collection with the preliminary scale and is considered a form of high-speed intuitive decision making (Scott & Bruce, 1995).

The five DMS are not mutually exclusive. While the rational and intuitive styles resemble the rational/intuitive dichotomy in the REI, the other three styles are less understood from a theoretical standpoint (Scott & Bruce, 1995; Thunholm, 2004). Correlations with measures of locus of control indicate that the rational decision makers correlated strongly with an internal locus of control, (i.e., that they believed in having the power to control outcomes), whereas other decision-making styles tended towards an external orientation (i.e., believing that outcomes are not in one's hands) (Scott & Bruce, 1995). The authors suggest that this may, in part, be due to a lack of confidence in one's decision-making capabilities in the case of the avoidant style, or a preference to delegate responsibility to others in the case of dependent decision making (Scott & Bruce, 1995).

Compared to the REI and REI-40 discussed in the previous section, there is less validation research available for the GDMSI. There are several studies, however, which confirm the five-factor structure, provide evidence of scale properties, or validate the GDMSI against other established measures of personality or behaviour (e.g., Baiocco et al., 2009; Bruine de Bruin et al., 2007; Gambetti et al., 2008; Loo,

2000; Spicer & Sadler-Smith, 2005; Thunholm, 2004). Spicer and Sadler-Smith (2005), for instance, were able to confirm the five-factor solution proposed by Scott and Bruce (1995) by both, confirmatory as well as exploratory factor analysis.

The GDMSI was also translated and confirmed in other languages (Gambetti et al., 2008) and specific styles were found to be predictive of performance on experimental decision tasks as well as real-life behaviours. Bruine de Bruin et al. (2009) found evidence that the rational style correlated with good decision making and high decision satisfaction, while the avoidant and spontaneous style correlated negatively with these two variables. In a study of predictive validity, Curşeu and Schruijer (2012) found that the rational style predicted decisiveness while the avoidant style predicted indecisiveness. In a study of circadian preferences, Tonetti et al. (2016) found evidence for the avoidant and spontaneous styles to be related to a preference for eveningness in young adults. Thunholm (2008) showed that cortisol levels correlated positively with the avoidant style and negatively with the spontaneous style. A final example of how the GDMSI was used is Fischer et al.'s study (2015), who successfully adapted the original questionnaire into a questionnaire of patient decision making.

A drawback of the GDMSI is that despite a body of publications and successful replications of scale structure, there appears to be no consensus yet within the field as to what decision-making styles *are*. Spicer and Sadler-Smith (2005) conclude in reference to Curry's (1983, in Spicer & Sadler-Smith, 2005) "onion" model of individual differences, which locates the most stable attributes such as personality and cognitive style in the centre and more malleable expressions of personality at the outer layers, that decision making styles could be "a surface manifestation of more deep seated personality constructs" (p. 146). Thunholm (2004) comes to a different

conclusion. He correlated the GDMSI with scales of self-esteem and self-regulation and concludes that decision-making styles are not merely a habit or propensity, but also depend on cognitive abilities (he names information processing, self-evaluation and self-regulation) and thus might be more similar to a trait and less flexible than habits which can be shaped by intervention if needed.

2.6.4 Judgement and Decision Research in Educational Measurement

In relation to rater cognition, there have been some attempts in the fields of educational measurement and language testing to engage with JDM research. Suto and Greateorex (2008), for example, mapped the strategies involved in rating mathematics and business studies items to the level of processing required (System 1, System 2, or a combination of the two). They were able to identify profiles for each of the two subjects which in turn could be used to more effectively distribute marking load among examiners and improve examiner training.

Crisp (2012) investigated the judgement processes involved in the grading of high-stakes project work as it was practiced by teachers in the context of the General Certificate of Secondary Education (GCSE) in the UK. Using verbal protocols, Crisp interviewed class teachers ($N = 13$, from three subjects), teachers in charge of school-internal moderation of the grades ($N = 3$, one per subject), and professional raters who each produced verbal protocols while assessing – or reviewing the assessment of – a set of projects ($N = 12$ per subject). Six rating categories emerged from the verbal protocol analysis: (1) Planning and orientating processes; (2) reading and understanding processes; (3) comments on task realization; (4) social and emotional reactions; (5) concurrent evaluation; and, (6) overall evaluation and score consideration. The findings provide an inventory of processes and behaviours

involved in the judgement of projects across different subjects. In the discussion, Crisp draws on Sadler's (1989) categorization of theories of judgement; the *analytic* approach, which assumes that relevant features of performance are noticed and weighted to make a judgement, and the *configurational* approach, which puts forward that a judge forms a mental image of a performance as a whole before making a judgement. Crisp found that some teachers might rely more heavily on the analytical approach than others. There was also evidence of some teachers balancing a kind of pattern recognition with a goal-oriented analytic approach. Moderators tended more towards the configurational approach which "may well reflect their greater experience and training" (p. 18). In light of these results, Crisp concludes "that neither (...) model of judgment can adequately represent what is involved in the assessment of projects" (2012, p. 17) and can only speculate as to why some teachers might rely more heavily on an analytic process, while others embrace a more configurational approach in spite of the rating mechanisms put into place.

Within the area of language testing, Baker's (2012) exploratory study tried to establish a connection between decision-making styles, rater cognition and rating scores. Baker selected six raters from a high-stakes writing examination and collected rating data (MFRM, central tendency and range of scores), observational data (deferred scores), and self-report data (write-aloud notes, General Decision-Making Style Inventory by Scott & Bruce, 1995). Even within this small sample of raters, Baker was able to identify distinct decision-making profiles and rating behaviours; for example, the underuse of failing bands or deferred rating decisions in avoidant decision-makers. The small sample, however, limited the generalizability of the findings. The discrepancies that Baker also addressed in the discussion of the results are related to self-report bias, the nature of the different sources of data, and the mismatch of MFRM

score patterns and decision-making style. Possibly because raters were at a rating session when participating in the study, they did not endorse the less socially desirable decision-making styles such as the avoidant or dependent style. Baker also hypothesized that lenient raters would tend toward the avoidant DMS which could not be confirmed for the two most lenient raters in the sample. Despite the mixed and inconclusive findings, Baker's study suggests the need to further investigate individual differences among raters as they may help explain some of the variability observed in the numerous rater cognition studies. Furthermore, a more comprehensive understanding of the nature of individual differences in rater cognition may help in creating materials and training procedures that are more tailored towards the individual needs of the participants.

2.7 Variables and Research Questions

The proposed project seeks to explore the nature of the rating process in the context of the Austrian Matura EFL speaking exam. It will do so by focussing on a group of novice raters. Specifically, it will investigate (1) the consistency and accuracy of novice raters in using the Austrian rating scale for speaking, as well as their perceptions of the scale; (2) the relationship between rater consistency/accuracy and rater attributes previously not considered in the context of rating speaking; and (3) profiles of rating behaviour in case studies selected to represent high- and low-accuracy raters.

This aim is broken down into sets of research questions and sub-research questions which will be addressed in three results chapters.

Chapter 4

1. How does a group of novice raters use the analytic rating scale currently in use for assessing speaking performances in the Austrian Matura examination?

1.1 How consistent are novice raters when using the rating scale? And does consistency vary according to scale criteria?

1.2 How accurate are novice raters when using the scale? And does accuracy vary according to scale criteria?

1.3 Are novice raters comfortable with applying the scale and are they confident in their rating decisions? Does this vary according to scale criteria?

Chapter 5

2. When novice raters assess speaking, are rating quality and rating behaviour metrics related to cognitive attributes, preferred processing mode or decision-making styles?

Chapter 6

3. To what extent do case studies of accurate and inaccurate raters reveal differences in rating behaviour and influences on rater behaviour?

3 Methodology

This chapter describes the methodology for this study. Sections 3.1 and 3.2 provide a rationale for the approach chosen for this study and describe the research design. Sections 3.3 to 3.8 provide details regarding the research context (3.3), participants (3.4), materials selected and developed (3.5), instruments (3.6), procedures (3.7) and methods of analysis (3.8) for each of the three sub-studies.

3.1 Mixed Methods in Rater Cognition Research

The main motivation behind this research project was to investigate rater cognition in novice raters when assessing speaking and to explore the potential role of attributes such as preferred cognitive style, cognitive capacity or decision-making style on the rating process and rating outcomes.

One characteristic of many rater cognition studies is that of comparing and contrasting different groups of raters or individual raters. This acknowledges the fact that raters are individuals who bring a particular set of attributes, skills, and strategies to the rating task, and that some of these features influence the rating processes and outcomes. As discussed in Section 2.5.1 (p. 48), studies have typically investigated group differences along variables such as rater experience (e.g., Attali, 2015; Barkaoui, 2010a; 2010b; 2011; Davis, 2008; Wolfe et al., 1998), first language (e.g., Kim, 2009; Y. Zhang & Elder, 2011; 2013), accent familiarity (e.g., Huang, 2013; Huang et al., 2016; Winke & Gass, 2012), or rating expertise based on rating quality metrics (e.g., Baker, 2012; Davis, 2008; J. Zhang, 2016; Wolfe, 2006). Baker (2012) may be among the first attempts so far to also compare raters and the rating processes

they might employ along a psychological attribute, their decision making style profile (Scott & Bruce, 1995).

A characteristic of current rater cognition research is a shift away from purely qualitative or quantitative research designs. In 2012, Wolfe and McVay criticised the fact that the majority of research into rater cognition falls into two separate strands: (1) research focused on rater characteristics and behaviour in order to provide information on how to facilitate quality rating, and (2) research into statistical indices and how they help identify problematic rating patterns (Wolfe & McVay, 2012). Both approaches come with a set of strengths and weaknesses. While the first, qualitative strand has the potential to offer insights into individual decision-making methods it “suffers [from] loosely-defined measures of 'rating quality' or no measure at all” (Wolfe & McVay, 2012, p. 36). The quantitative strand, on the other hand, is useful in that it identifies issues with a rater cohort or individual raters; however, it fails to explore the underlying reasons for conspicuous rating patterns (see also Yan, 2014). Recent research into rater cognition seems to have heeded Wolfe and McVay’s (2012) call for combining both strands and there has been a considerable increase in studies that employ a mixed methods approach to identify the sources of rater variability and help explain why raters behave in a certain way (e.g., Baker, 2012; Davis, 2008; Lumley, 2005; Yan, 2014; J. Zhang, 2016). The emphasis tends to remain on quantitative methods in many of these studies as the analysis of the qualitative data is labour intensive (Davis, 2015; 2008) or contributes to a considerably smaller part of the argument (Yan, 2014).

This trend runs parallel to a general development in the social sciences where the mixed-methods research (MMR) approach has been accepted as a viable third

methodological paradigm in order to investigate complex social phenomena (Dörnyei, 2007). MMR, similar to the argument-based approach discussed in Chapter 2, is grounded in pragmatic philosophy. It is argued that combining different methods of data collection and analysis improves the validity and, ultimately, the generalizability of research results (Creswell & Plano Clark, 2018). The MMR approach acknowledges that each strand, qualitative and quantitative has the potential to contribute unique insights. Hence, the purposes of MMR may be 1) to triangulate data in order to cross-validate insights from various angles, or 2) to “elaborate, clarify and explain the results from one method with the results from another method” (Jang et al., 2014, p. 129).

Since MMR found its general acceptance towards the end of the twentieth century, a number of authors have suggested various research designs (Creswell & Plano Clark, 2017) and methods of formally describing these designs (Ivankova, 2015 in Ivankova & Greer, 2015). One option is to concurrently collect quantitative and qualitative data to merge it during interpretation. In this case one type of data complements the other. Furthermore, there are two sequential design options. Either quantitative data informs a qualitative investigation to elaborate, explain or confirm the findings from the quantitative stage (sequential Quan to Qual MMR design), or qualitative data helps form categories or themes to then be operationalized by quantitative instruments (sequential Qual to Quan MMR design).

A particular challenge of the MMR paradigm is that the researcher needs to navigate and consolidate different philosophical stances within one research project (Creswell & Plano Clark, 2018; Ivankova & Greer, 2015; Mackey & Gass, 2015; Teddlie & Tashakkori, 2009). On the one hand, the development and application of empirical

instruments and analysis of quantitative data is grounded in a (post)positivist worldview. The gathering and analysis of qualitative data, on the other hand, requires adopting a constructivist stance in that it assumes that the researchers accept a multitude of subjective views from different participants' perspectives to form broad understandings of the phenomenon under investigation (Denzin, 2012). Depending on the research design, a researcher might be required to switch between paradigms in different phases of the project (i.e., sequential designs), or even consider both at the same time (i.e., concurrent designs) (e.g., Creswell & Plano Clark, 2018; Ivankova & Greer, 2015).

The MMR paradigm is also highly compatible with an argument-based approach to test validation as the argument-based approach offers a framework to integrate data from qualitative and quantitative methods (Xi, 2008). Kane (1992), for example, argues that “parallel lines of evidence” (p. 528), where evidence gathered by various sources provide support for a particular claim and make a practical argument more powerful and convincing. Similarly, there have been repeated calls from within the field of language testing to support conclusions about test validity with more than one source (e.g., Bachman, 1990; Wolfe & McVay, 2012). As many sources of data are combined, validation research and MMR-based studies run the risk of amassing great quantities of data without establishing or supporting a clear conclusion. Therefore, it is important to gear these research efforts towards clearly defined goals.

3.2 Selected Approach

The research design for this study followed the principles of MMR and combined several sources of data through a series of three linked studies (see Table 3.1).

Study 1 investigated how a group of novice raters handled the assessment of spoken performances. This study included quantitative as well as qualitative data. Rating quality (severity, accuracy, fit and bias) was estimated by analysing a set of ratings provided by the participants to a common set of performances (the quantitative component). Self-report Likert items and open-ended items were analysed to capture the experiential dimension of rating in general and of using the rating scale (the qualitative component).

Study 2 explored whether cognitive attributes, cognitive preferences or preferred decision-making style influenced rating quality or rating behaviour. This study was exclusively quantitative and consisted of a series of correlation and regression analyses. The variables were the rating quality measures taken from Study 1 as well as time stamp data, cognitive test scores, questionnaire data and Likert-scale items on perceived difficulty and confidence when rating the speaking performances.

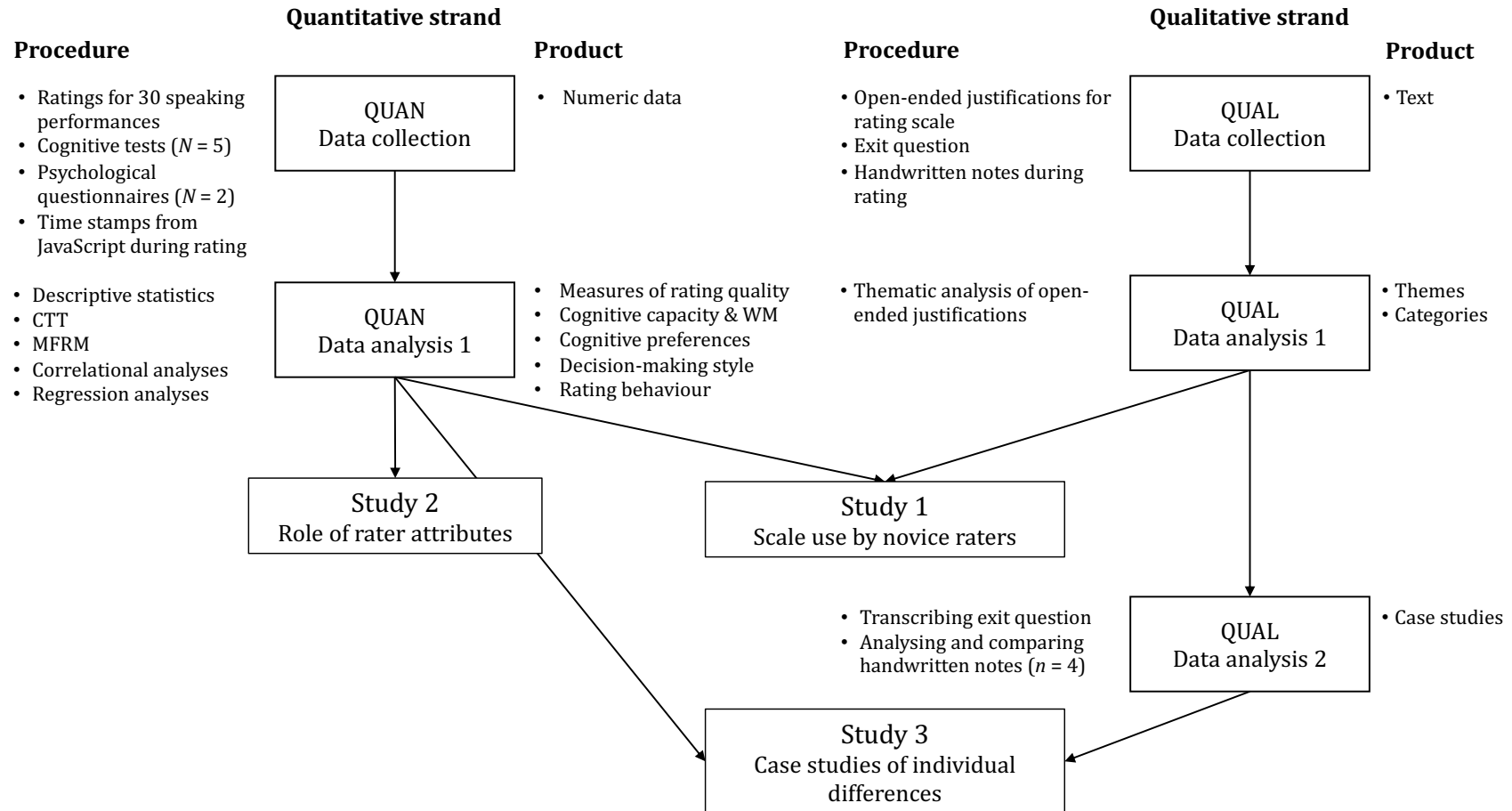
Study 3 took a case study approach and sought to explain individual differences in rater behaviour. Four substantially interesting raters were selected based on their accuracy (Study 1), but also other selected variables identified in Study 2. Quantitative data from the first two studies was revisited and drawn on to provide a narrative insight into the rating experience of the individual rater. Each case was also augmented with additional qualitative data – observations from raters' handwritten notes made during the rating process, and comments from an open-ended exit question – providing additional insight into each rater's rating experience.

Table 3.1 Research design of three linked studies

Study	Research questions	Data	Analysis
Study 1 Scale use by novice raters (Chapter 4)	How does a group of novice raters use the Austrian rating scale? 1.1 How consistent are raters? 1.2 How accurate are raters? 1.3 Are raters comfortable with applying the scale and are they confident in their decisions?	Scores Perception data (self-report) General confidence Perceived difficulty of rating as a whole Perceived difficulty of using criteria	CTT Interrater reliability Accuracy MFRM analysis Rating scale model Rater accuracy model Thematic analysis
Study 2 Role of rater attributes (Chapter 5)	Are rating quality and rating behaviour metrics related to cognitive attributes, preferred processing mode, or decision-making style?	Rating quality measures (Study 1) Rating behaviour metrics (time stamps) Cognitive test scores Questionnaire data REI-40 (preferred cognitive mode) GDMSI (decision-making style)	Descriptive statistics Correlational analyses Regression analyses
Study 3 Case studies (Chapter 6)	How do individual raters with different accuracy profiles differ in terms of their experience during the rating task?	Results from Study 1 and 2 Handwritten notes from rating session Responses to exit question	Selection and compilation of 4 case studies

The overarching MMR approach of this study is thus an explanatory sequential quant-qual design, where a quantitative phase (Study 1 and Study 2) is followed by a second, more qualitatively oriented phase (Study 3). This model lends itself to forming groups or identifying individuals based on a quantitative variable and then investigating group differences more thoroughly in a subsequent qualitative phase (Creswell & Plano Clark, 2018) and aligns with the recommendations for rater cognition research expressed by Wolfe and McVay (2012). Creswell and Plano Clark (2018) suggest that only one quantitative variable is of interest during the first stage, but the approach for this study is more flexible. The three individual studies vary in their orientation in that Study 1 presents quantitative data alongside qualitative data, Study 2 is strictly quantitative, and Study 3 uses results from the previous studies as a focal point for further qualitative investigation. This adaptation is attributed to the complexity of the phenomenon under investigation as it would be against the purposes of this study to consider rating quality and behaviour only in terms of a single variable. Figure 3.1 provides an overview of how data from various sources were combined and integrated in each of the three studies.

Figure 3.1 Overview of quantitative and qualitative data integration



3.3 Research Context

In Austria, there are several routes by which to qualify for enrolment in tertiary education. Most of the Austrian university or college students – roughly sixty percent – have passed the *Reifeprüfung*, or *Matura* (BMBF, Bundesministerium für Wissenschaft, Forschung und Wirtschaft, 2014). The *Matura* consists of three components. The first component is a research paper on a subject of the student's choice between 40,000 to 60,000 characters. After submitting the paper, students then also present their work during one of their final oral examinations. The second component is the written examination which comprises the compulsory subjects Mathematics, German as a first language, and one foreign language – often English, French, Italian or Spanish. Each written exam is between four to five hours. The third and final part of the *Matura* is the oral examination. Students are free to choose subjects for the oral examination according to their interests unless they fail one of the written examinations. In that case taking an additional oral exam in the failed subject is compulsory to improve the grade and still obtain the certification. The repercussion of this set up is that students can opt to either take the written or oral examination, or both, for the subject English. In 2018, for instance, only 25,488 of 37,467 students took the written examination in English (15,666 general upper-secondary and 8,822 vocational upper-secondary) (Bönisch et al., 2020). There are no public records of candidate numbers for the non-standardised oral examination.

In its current revised form, the written component of the foreign languages *Matura* includes three standardised parts – Writing, Listening and Reading – for all school types and an additional grammar and vocabulary component called Language in Use

for general upper-secondary schools. The targeted level for the first foreign language subject, which is English for most students in Austria, is the B2 level as described in the Common European Framework of Reference (CEFR, Council of Europe, 2001).

The English speaking examination itself takes 15 minutes and consists of a monologue (long turn) (about five minutes) and a conversation with the teacher or with another fellow student (about ten minutes). Students are granted at least 15 minutes preparation time before the examination. The tasks are created by the class teacher according to a set of sample tasks and guidelines first published by the Ministry of Education in 2013 (BMBF, 2013). The tasks are meant to be a guideline for teachers and the collection was created by a group of highly experienced classroom teachers whose names can be found in the document. Since 2020, the Österreichische Sprachen-Kompetenz-Zentrum (ÖSZ, 2020) has published a new set of sample tasks. For this set, the collaborative development process involved experienced teachers and language testers, and the tasks were tried with students before a final review.

The task for the monologue, or individual long turn, often consists of a picture prompt like the one presented in the ÖSZ sample task (see Figure 3.2) and must include three bullet points. Each bullet point is designed to elicit a particular language function which candidates must address. The discussion task, or paired activity, is always set in a certain context (e.g., reacting to an advertisement from a British news site, see sample task in Figure 3.3) and usually contains a list of options that candidates need to discuss and agree on to fulfil the task demands.

Figure 3.2 Sample individual long turn (based on ÖSZ, 2020, p. 30)

PICTURE OF TASK REMOVED FOR COPYRIGHT REASONS

Description: Individual long turn

Give a five-minute talk on the topic of immigration in which you

- compare the ideas these pictures stand for
- analyse the major reason why people have come to seek refuge in Austria,
- make suggestions how Austria could help young refugees integrate.

Figure 3.3 Sample paired activity task (based on ÖSZ, 2020, p. 31)

PICTURE OF TASK REMOVED FOR COPYRIGHT REASONS

Description: Paired activity

The British online news site *THE LOCAL* (www.thelocal.uk) is planning to run a series on Austrians abroad. One podcast is going to deal with the reasons why young Austrians emigrate, so they have asked people to send in two ideas.

You and your partner have decided to send in two ideas. You have 10 minutes to discuss the following options and decide which two you are going to suggest:

- career options
- education
- love
- way of life
- adventure

In order to help improve the transparency of grades attained from the English speaking test, the Ministry of Education implemented the mandatory use of centrally distributed analytic and holistic rating scales (see Appendix B.1 and B.2; the structure of the analytic rating scale is discussed in Section 3.6.1). The development of the scales was commissioned to a team of language assessment specialists from the University of Innsbruck and an external consultant. Several aspects of the scale development process are particularly noteworthy. Right from the outset, the project involved ten highly experienced and trained language teachers from various Austrian regions and with various languages as their teaching subjects. The scale development process started by gleaning relevant scales from the CEFR, but also involved comparing the emerging version of the scale and its descriptors against actual student performances. The process, thus, combined an ‘intuitive’ as well as an ‘empirical’ approach to scale design (Council of Europe, 2001), or what Fulcher et al. (2011) labelled as ‘measurement-driven’ and ‘performance-driven’. Finally, the scales were developed for use across several languages and for the A2, B1 and B2 level of the CEFR. This meant that the scales and descriptors were discussed in plenary sessions as well as refined in smaller team meetings (for more details see Holzknecht et al., 2018).

After the analytic rating scales had been finalised, the team also created holistic rating scales (see Appendix B.2). The holistic scale basically included one descriptor that was considered particularly useful from each criterion. The rating scales were published along with a very brief description and instructions on how to use them during the examination (BMBF, 2012). Austrian teachers do not require additional qualifications to run or assess examinations and they may or may not be introduced to the scales during their pre-service studies. Moreover, teachers are not obliged to seek out training opportunities with the scale. Right after the new scales had been

introduced, regular seminars (between half a day and a full day) were offered across Austria. However, without any public records it is unclear how many teachers received formal training at these seminars and how many familiarized themselves with the scale via online materials or in exchange with a colleague.

3.4 Participants

Three sets of participants took part in this study: (1) the speakers who produced the speaking performances, (2) the expert judges who rated the speaking samples to create reference scores, and (3) the teacher students who participated in the rating experiment and who provided the main data set.

3.4.1 Speakers

As there is only a handful of recorded sample performances for the English *Matura*, a set of performances was specifically recorded and rated for this study. 27 speakers from two Tyrolean schools and 14 early-stage university students were recruited to sample a broad range of ability levels. Recordings took place in spring 2016. The 17 students from the Gymnasium Reutte (a rural school) and Akademisches Gymnasium Innsbruck (an urban school) were in their final year and close to their final examinations. All speakers (or their parents) were informed about the purposes of the test simulation and participated on a voluntary basis after signing a consent form (Appendix N). There were only six male speakers. This low proportion is partly due to the student population at university. However, it is also possible that male students were more reluctant to volunteer for a simulated speaking examination.

After the recording, the performances from these initial 41 speakers (6 male, 35 female) were rated by experts. From this set of 41 speakers, 30 performances were selected to go into the final pool for the experiment (see section 3.5.2 Reference Scores for details on the selection process).

3.4.2 Expert Raters

All video-recorded performances were first assessed by a group of six testing experts and teachers via the online survey platform Qualtrics. All raters had postgraduate degrees in language testing and varying degrees of experience in rating speaking and language teaching. Two experts were involved in creating the analytic scales used in this study, a process which included the rating of performances. Two other experts were involved in running rater training sessions for this scale. The last two experts were active teachers at university level English courses and were also involved in scale development workshops. All raters were very familiar with the rating scale.

3.4.3 Raters

A cohort of 39 novice raters were recruited for the study. There were several reasons why this study focused on novice raters. First, this group was expected to display greater homogeneity along the factors identified earlier in this dissertation (experience, expertise, and teaching experience). It was hoped that creating a more homogeneous participant cohort and controlling for influential factors would maximise the chance of isolating the effect of lesser-known factors explored in this study (i.e., cognitive factors, cognitive preferences, and decision-making styles). Second, novice raters' cognition was expected to be more sensitive to the urgencies of rating speaking in real time because they have not yet "adopt(ed) a particular approach

to the rating task” (Eckes, 2015, p. 31) and are untainted by what Lumley (2005) refers to as the “institutional level” (p. 289) (i.e., the constraints and goals of assessment in a social context). Finally, the perception and use of the rating scale by novice raters has the potential to provide valuable insights into scale functionality. Referring to Alderson’s (1991) definition of assessor-oriented scales, Fulcher (2003) argues that rating scales need to be worded in a way that promotes fast processing during the exam situation and produce acceptable reliability even without providing ample training, so that reliability does not become an artefact of training rather than a scale property. These arguments are particularly salient in the case of the Austrian *Matura*, where the use of the rating scale is stipulated by law, but not all teachers are likely to have attended or even sought face-to-face training prior to implementing the rating scale.³

The novice raters were recruited from the then current student body of the School of Education at the University of Innsbruck. They were undergraduate students and pre-service Austrian teachers of English. The minimum language requirement to enter teacher training programs is either passing the *Matura* or, for some students who passed the *Matura* in another language than English, a positive grade in their last year of studies. Both, the syllabus and the examination require reaching the B2 level of the CEFR. To minimize possible differences in English language proficiency among the participants, the recruitment process targeted students toward the middle and end of their studies, i.e., after the fifth semester of studies. It was hoped that at this point of their training, participants’ English proficiency levels would be more homogenous and

³ Unfortunately, there are no centralized records on the number of trainings that were offered in each region, or in Austria as a whole, and estimated number of attendees.

solidly at the upper end of the B2 level and above as is specified in the course syllabus.⁴

It is important to highlight two further particularities of this sample. First, Innsbruck University trains teachers for Austrian and Italian secondary upper-level schools, for the autonomous Italian province of South Tyrol. Second, Austrian study programmes tend to be much less streamlined than in other countries. Students do not progress in cohorts, instead they might opt to or, in case of overcrowded courses, might have to study more slowly in one subject area and move ahead faster in another. Furthermore, many Austrian students work part time during their studies. With a minimum length of eleven semesters and no stipulated maximum length it is, therefore, not uncommon to find students that have already studied for six years (i.e., twelve semesters) or longer. These factors make the recruitment of a fully homogenous sample in terms of length of studies, age and nationality challenging. Thus, the main inclusion criterion was being an EFL teacher student with at least five semesters of studies.

Participants were recruited via a two-stage combination of convenience and snowball sampling (e.g., Miyahara, 2019; Wagner, 2015). First, all students taking the compulsory Language Testing and Assessment course in the winter and summer terms were informed about the possibility of participation. This course is part of the last module of their subject Didactics. The topic of the study was presented briefly during the course and participants could then sign up. To meet the target of 40 participants, students who had been enrolled in this course over the previous two years were

⁴ The entry requirement for any teacher training is a Matura, which means obtaining a positive grade in the final school year or, much more common, passing the standardized examination in English. Both pathways are set at the B2 level.

contacted via an email and a targeted Facebook post on an English student group site, as well as via direct contacts to former students.

Participants received 70€ compensation as they were required to come to the department on three to four separate occasions, requiring a commitment of five hours or more. During the promotion of the project, the students were told that their participation would be a good investment in their professional development as they would get a two-hour rater training (see Section 3.7.1 for details) at the beginning of data collection as well as the actual experience of rating 30 performances and feedback on their rater behaviour after the end of data collection.

The recruitment and data collection were carried out in four parallel cohorts from October 2017 until April 2018. From an estimated number of about 350 eligible students that were contacted, 39 (4 male, 35 female; mean age 24.28 years, mean length of studies 8.64 semesters) volunteered to participate. The gender imbalance in the sample is representative of the typical student cohort in that there is a larger proportion of women in this programme (see

Table 3.2).

In the following chapters and to make the description clearer, this third participant group of English pre-service teacher students will be referred to as *raters* or *participants*. Whenever needed, the first participant group will be referred to as *speakers* and the second group as *expert raters*.

Table 3.2 Overview of raters

Name*	Sex	Age	Semester
Alice	F	23	9
Amber	f	24	8
Amy	f	23	7
Betty	f	22	6
Chloe	f	25	7
Daisy	f	24	8
Deborah	f	23	7
Doherty	f	22	6
Donnie	m	23	8
Eliza	f	22	6
Esme	f	26	11
Helen	f	27	13
Hendrix	m	26	8
Holly	f	22	6
Jennifer	f	23	6
Kimberly	f	21	6
Leila	f	23	7
Lexi	f	27	11
Lucy	f	24	7
Maddison	f	32	23
Maisie	f	23	7
Margarete	f	22	6
Mary	f	23	5
Megan	f	27	12
Nancy	f	22	7
North	f	23	8
Paige	f	29	15
Penelope	f	23	7
Phoebe	f	24	7
Rosie	f	27	12
Sally	f	23	10
Scarlet	f	23	7
Stormi	f	23	6
Susan	f	27	11
Tyler	m	29	11
Violet	f	24	8
Willow	f	23	7
Zachary	m	29	15
Zara	f	21	6
<i>M</i>		24.28	8.64
<i>Mdn</i>		23	7

Note. * Participants chose their pseudonyms for this study.

3.5 Creating and Selecting Performance Samples

3.5.1 Video Recording

As rating in a live test administration would not have been a feasible method to address the research questions, it was decided to use video-recorded performances. 41 individual speakers were initially recorded. Speaker participants were invited to an exam session at their school (26 school students from a rural and an urban upper-secondary school) or at the University of Innsbruck (15 students). The purpose behind the recordings was explained to them and all speakers (or their parents) provided consent prior to the simulation.

The simulated examinations followed the exam guidelines in that speakers would first do an individual long turn followed by a paired activity with another student. However, after a short break, speakers would do an additional individual long turn task⁵. Another departure of the simulated examination was that instead of the official 15 minutes preparation time, speakers were given three minutes to prepare the individual long turn. There were several arguments for shortening the preparation time: 1) the performances would capture more spontaneous speaking, 2) the effect of test preparation training would be levelled out, and 3) the task would be more challenging and make the differences between the speakers more visible. A shorter preparation period also had practical advantages when recording the sessions.

⁵ In the initial stages of this research project, it had not been decided yet, whether or not raters would only rate individual long turns. Once the design was set it became clear that focusing on the monologue would help narrow down the number of variables and also make the rating load for the participating raters more manageable.

All tasks were in line with the guidelines published by the Ministry (BMBF, 2013) to resemble typical tasks in use for the examination (see Appendix C for tasks). To standardize procedures, I acted as the only interlocutor in all simulated examinations. I was involved in the production of interlocutor guidelines and had trained teachers in interlocutor training sessions. Using local teachers as interlocutors appeared too big a risk as it was unclear how they would perform when simulating an assessment situation with their own students.

Six performances had to be discarded after the recording sessions due to technical problems (e.g., sound quality, storage problems, picture quality), which left a total of 76 speaking performances available for this study (including both long turn and interactive tasks).

The individual video files were edited and compressed to .mp4 (720p) format via iMovie (ver. 10.1.8; Apple 2001-2017) and then further compressed and optimised for video streaming (codec H.264) via Handbrake (ver. 0.10.5; HandBrake Developers 2003-2016). The videos were uploaded onto a private YouTube channel using the unlisted mode which ensures the videos can only be accessed via an URL link. Finally, the videos were embedded into a Qualtrics survey (Qualtrics, Provo, UT), which was then used to distribute the performances to the expert raters and creating the actual rating survey.

3.5.2 Reference Scores

Employing a rotated rating plan (Linacre, 1994) or incomplete spiral design (Eckes, 2015), each of the initial 76 performances was rated by at least three of the six expert raters (see Appendix D for rating plan). To increase connectedness between the

observations, twelve performances were rated by all raters; these were six speakers out of the 76 recorded for this study and another six performances that had been benchmarked for another project. The rating plan was piloted by using authentic rating data from a rater training workshop. A Many-Facet Rasch Measurement (MFRM) analysis was carried out using FACETS (Linacre & Wright, 1992-96) and following a procedure described by Eckes (2015) to yield adjusted scores based on the fair average ratings for each performance and on each criterion. This procedure remediates the relative severity or leniency of any expert rating of a performance and identifies misfitting or problematic performances. Table 3.3 reports on the rater measurement estimates based on the first step of the fair score procedure whereby the rater facet remains non-centred before separate analyses are carried out to establish the fair score of a performance on each criterion (see Eckes, 2015). The MS fit indices for all six expert raters were within the range recommended for rating scale data (0.6-1.4) (e.g., Bond & Fox, p. 273) and were found to be useful for the fair score procedure.

Table 3.3 Rater measurement report

Rater	Obs. Score	<i>M</i>	Fair <i>M</i>	Measure	<i>SE</i>	<i>MS_w</i>	<i>t_w</i>	<i>MS_u</i>	<i>t_u</i>
2	1201	7.32	7.33	-0.45	.09	0.90	-0.8	0.88	-0.9
3	1270	7.74	7.79	-1.03	.09	1.05	0.4	1.05	0.4
6	1329	8.10	8.19	-1.52	.09	1.06	0.5	1.27	1.8
4	1363	8.31	8.43	-1.81	.09	1.09	0.8	1.29	1.9
1	1419	8.65	8.85	-2.32	.10	.90	-0.8	0.79	-1.4
5	1422	8.67	8.88	-2.37	.10	.86	-1.3	1.11	0.7

Note. *MS_w* = mean-square infit statistic. *t_w* = standardized infit statistic. *MS_u* = mean-square outfit statistic. *t_u* = standardized outfit statistic. Estimates based on 164 observations per criterion.

The FACETS analysis revealed that the performances spanned the entire B2 level as reflected by the Austrian rating scale, with performances barely passing the minimum requirements to performances at an advanced B2 level. The analysis showed that one

of the two tasks (“Food waste”) used to elicit the performances discriminated more successfully than the other (“Enjoying music”) (Appendix C). Thus, out of the entire corpus of 76 performances, 30 performances from 30 speakers on the task “Food waste” were selected for the main data collection. To select the final set, all performances with significant infit or outfit statistics or considerable corrections after the fair score procedure were generally excluded as there may have been something about this performance that made it more difficult to rate. After this step, aspects such as gender of speaker, location of recording (school or university) and ability level were considered to include a broad, but balanced selection of the performances in the experiment. The logit scores from the MFRM analysis were then used to create two roughly similar sets of 15 performances, balanced in terms of average logit measures and mean fair averages. These two sets formed the materials to be used for the two rating sessions in the main study (see Table 3.4; see also Appendix E for details on each set of performances). This ensured that both rating sessions would include speakers of a similar range of speaking ability.

Table 3.4 Mean fair average scores in Session 1 and 2

Criterion	Session 1	Session 2
TA	8.21	8.05
FLIN	8.00	7.83
RSL	8.10	7.93
ASL	7.99	7.83

Note. TA = task achievement. FLIN = fluency and interaction. RSL = range of spoken language. ASL = accuracy of spoken language.

3.6 Instruments

3.6.1 Austrian Matura B2 Rating Scale

The central instrument for this study is the Austrian CEFR-linked assessment scale at B2 level which was first published by the Ministry of Education in 2012 (BMBF, 2012). The scale was developed as part of a larger project coordinated by a team at the University of Innsbruck. The goal was to produce analytic and holistic rating scales for the levels A2, B1 and B2 which could be used by Austrian teachers for foreign language teaching and assessment at the upper-secondary level. When the rating scales were finally published in 2012, teachers were obliged by law to also use them in the oral examination of the Matura. As they are the first centrally developed and distributed assessment scales in Austria, it was decided in the early development stages to adapt a basic scale structure that somewhat resembles the then also new rating scale for writing, allowing for easy arithmetic conversion into school grades (see Holzknicht et al., 2018).

The analytic rating scale for speaking at the B2 level consists of four equally weighted criteria (Task achievement [TA], Fluency and interaction [FLIN], Range of spoken language [RSL] and Accuracy of spoken language [ASL]; in that order) and is divided into eleven bands (0-10) (see Appendix B for full scale). *Task achievement* (TA) is the first criterion in the scale and contains descriptors on whether the candidates have addressed all aspects of the task, the quality of thematic development and how well they are able to support and sustain their opinions. The criterion *Fluency and interaction* (FLIN) combines descriptors on the perceived fluency of the performance and to what degree candidates can take turns and react flexibly in the context of the

task. The criterion *Range of spoken language* (RSL) describes the range of different structural and lexical means that are available to the candidates. Finally, the criterion *Accuracy of spoken language* (ASL) offers descriptors on lexical and grammatical accuracy and pronunciation. The criteria are equally weighted which naturally leads to an emphasis on linguistic competences over TA or FLIN as two out of four criteria solely focus on linguistic competencies. To make the scale more user friendly, the scale includes empty as well as described bands (Holzknecht et al., 2018). Six bands (0, 2, 4, 6, 8, and 10) include descriptors and the other five bands are left blank. To illustrate the focal points of each criterion more specifically, Table 3.5 presents the descriptors for the minimum pass level (Band 6).

The rating scales were published along with a brief description and instructions on how to use them during an examination (BMBF, 2012). According to this document, all four criteria need to be applied to both parts of the test, the individual long turn and the paired activity. To pass, a performance must meet the overall requirements described at band six. Thus, the scale was designed in such a way that descriptors from B1 and upper B1 were used and adapted below band six of the rating scale and descriptors from B2 and upper B2 scales in the CEFR were used in the levels above band six. When rating analytically, all four criteria need to be applied to both parts of the test, the individual long turn and the paired activity. The teacher who assesses using the analytic rating scale combines their rating decisions with the holistic rating from their colleague and must form a final overall grade for the candidate.

Table 3.5 Descriptors for minimum pass level

Band	TA	FLIN*	RSL	ASL
6	(1) Most aspects of the task addressed and sufficiently expanded (2) Clear, detailed descriptions and presentations, expanding and supporting ideas with subsidiary points (3) Accounts for and sustains opinions by providing relevant support	(1) Fluent and spontaneous performance, causing no strain on the listener (2) Effective turntaking, not always elegant (3) Adjusts to changes of direction in conversation (4) Produces stretches of language with a fairly even tempo; few noticeably long pauses	(1) Sufficient range of language for the task, some restriction (2) Good range of vocabulary for the task, varies formulation to avoid frequent repetition (3) Can use circumlocution and paraphrase (4) Uses some complex structures	(1) Lexical accuracy generally high, mistakes do not hinder communication (2) Grammatical control relatively high; any mistakes do not cause misunderstanding (3) Can correct slips and errors if she/he becomes conscious of them (4) Clear, natural pronunciation and intonation

Note. TA = Task Achievement. FLIN = Fluency and Interaction. RSL = Range of Spoken Language. ASL = Accuracy of Spoken Language. *Raters were asked to disregard descriptors 2 and 3 in the FLIN scale as these descriptors pertain to the paired activity.

The analytic scale (Appendix B.1) is the only scale that was used for this study. Participants were given the scale in its original form. As two out of four descriptors from the FLIN scale only pertain to the paired activity (which is the second part of the examination and not included in this study), participants were shown and explained during the scale familiarisation session which descriptors would not apply to the performances they would see in the experiment.

3.6.2 Online Rating Platform

The speaking performances were presented via the online survey tool Qualtrics (Provo, UT). Qualtrics was also used to record the participants' rating decisions, their responses to self-report items regarding the rating process (see Section 3.6.4), and to capture time stamp data in the background. Two rating sessions with 15 performances each were set up in Qualtrics (see Section 3.5.2 for details on how rating sessions were compiled). Each performance was presented on a separate page with a grid representing the criteria and bands of the analytic rating scale just below the embedded video (Figure 3.4). Each rating session was split in half with eight performances in the first and seven performances in the second half. To avoid order effects, performances were randomised within each half. This meant that every participant would see the same performances in the first or second half of their rating sessions, but the performances were presented in a randomised order.

Through the embedded data feature integrated in Qualtrics (Provo, UT), additional time stamp data was recorded in the background. JavaScript was embedded manually in the script of each single survey element which enables the logging of time stamps for mouse events. This data offers additional insight as it traces the sequence in which

the analytic rating decisions were taken, the relative timing of the rating decisions, and whether or how often any rating decisions were revised before submission.

Figure 3.4 Screenshot of rating as presented in Qualtrics

View the video below and rate the performance:

P06 02 Foodwaste

	1	2	3	4	5	6	7	8	9	10
Task achievement	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fluency	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Lexical and structural range	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Lexical and structural accuracy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

In addition to the data recorded via Qualtrics, the participants were also given hardcopy rating forms that are like those used in the Austrian *Matura* (see Appendix F). These forms served two functions. First, in line with what raters are expected to be doing during a live examination, participants were instructed to use the forms for taking notes during the rating. Second, even though Qualtrics proved a reliable tool for this project, it was considered important to have hardcopy backup records of the rating decisions as they were the most salient data for this study.

3.6.3 Cognitive Tests

A range of instruments were employed to address Research Question 2 (Study 2), which focused on cognitive abilities, preferred processing mode and decision-making styles. In terms of cognitive abilities, based on the considerations discussed in the literature review above, it was decided to compile a cognitive test battery consisting of five measures. The three main basal functions of the central executive as described in Miyake et al. (2000) (i.e., shifting, updating and inhibition) were targeted with three established and validated tests:

- a) A Letters-numbers task, which captures the ability to shift attention between different tasks (Miyake et al., 2000; Rogers & Monsell, 1995),
- b) an adapted version of the Keep Track task to measure updating (Yntema, 1963),
- c) and a Stroop test to target inhibition (Stroop, 1935; as described in Miyake et al., 2000).

Second, an auditory forward digit span test targeted phonological short-term memory (e.g., Brunfaut & Révész, 2015; Indrarathne & Kormos, 2017). Third, the Trailmaking test was used as a measure of attention control (Isaacs & Trofimovich, 2010). The following section describes how each test was prepared for the data collection. Instructions and sample items for all tasks are provided in Appendix J.

Letters-numbers task. The Letters-numbers task was adapted from an experiment in PsyToolkit which is an online platform that allows customisation and off-line delivery of a range of cognitive tests (<http://www.psychtoolkit.org>). Each stimulus consisted of a letter-number combination (e.g., G6) presented in one of four quadrants. If the combination appears in the top half of the quadrant, participants must indicate whether the letter is a consonant or a vowel. If the combination appears in the lower half of the quadrant, the participants need to identify whether the number is odd or even. Thus, the task switches depending on the location of the stimuli. Participants react via

pressing two keys: *B* for consonants and odd numbers, and *N* for vowels and even numbers. The instructions were in English and followed up on in German. Participants also completed training items for all three trial types before starting with the test proper.

First, participants completed 32 trials with stimuli presented in the top two quadrants, which was the letter task. Next, they completed 32 trials with the stimuli presented in the lower two quadrants where they had to respond to the number, also called the number task. Finally, 128 trials were presented in which the position of the stimulus rotated clockwise, forcing participants to either repeat the same task as before (quadrants two and four), or switching between the letter task and the number task (quadrants one and three). The dependent measure was the shift cost in milliseconds between congruent and incongruent trials as calculated by the website.

Keep Track task. The Keep Track task was based on the description in Miyake et al. (2000) and presented via PowerPoint. This task required participants to identify and memorize particular words from a quickly presented sequence which requires flexibility in holding and updating information stored in working memory. First, the participants were familiarised with six target categories (*Tiere, Farben, Metalle, Länder, Währungen, Familie*) and the three elements within each category (e.g., *Vater, Schwester* and *Mutter* for the category *Familie*). For each trial, a series of 15 words was presented sequentially at a speed of 1500 milliseconds per word. Each sequence included two or three words from each category. At the beginning of the trial, the target categories were introduced. They remained visible throughout the trial at the bottom of each stimulus. For each trial, participants were instructed to watch the sequence of words and then write down those words that were presented last for each

of the target categories. After a training run with three target categories, the participants received feedback. The main test included three trials with four target categories and three trials with five target categories. The maximum score was the proportion of correctly recalled words out of the 27 items.

Stroop test. The Stroop test was designed and delivered via the PsyToolkit platform. In the test, participants must quickly and accurately identify the colour in which the name of a colour is printed. If the stimulus word is “red”, but it is displayed with a white font the correct answer is “white”. The task measures reaction delay between congruent tasks, where the word and the presentation colour match, and incongruent tasks, where the word and the font colour do not match. The Stroop test is language sensitive, and the instructions and stimuli were presented in German. By avoiding the L2 for this task, the participants’ L2 competence may not confound the actual measure of inhibition (e.g., Brunfaut & Révész, 2015). After receiving instructions, participants completed 72 trials with colour words printed in the same colour (e.g., ROT in red colour) and colour words printed in a different colour (e.g., BLAU printed in green). The trials were based on four colours (white, green, red, blue) and participants had to respond via pressing the initial letters for the German words (*w*, *g*, *r*, and *b*). The dependent measure was the interference time in milliseconds, or difference in reaction time, between congruent and incongruent trials as calculated by the website.

Forward Digit Span. The forward digit span was created and delivered via Microsoft PowerPoint. In this test, participants were required to recall correctly an aurally presented string of digits. This gauges the amount of information participants can store successfully in their working memory. A PowerPoint presentation was designed which presented slides at a rate of one digit per second and each slide contained one sound

file per number. As this test must be considered language sensitive, the stimuli were recorded and presented in German. After receiving instruction and practicing with two three-digit trials, the participants heard 14 increasingly longer sequences (i.e., seven pairs) ranging from three to nine digits (see also Brunfaut & Révész, 2015). Same as with the Keep Track task, participants recorded their answers on a response sheet which was compared to an answer key at the end of the test. The participants' digit span was the longest correct span out of all spans.

Trail Making task. The Trail Making task consisted of two separate parts, A and B. Participants used a pen to connect a set of numbers (Part A) or numbers and letters (Part B) in ascending sequence. Participants were instructed to not lift the pen from off the paper and to work as quickly as possible. Each part was preceded by a practice task. During Part A, participants were required to connect the numbers 1-25 (1-2-3-4-5 ...), and during the more demanding Part B, participants had to alternate between letters and numbers in an ascending order (1-A-2-B-3-C...). The dependent variable was the difference in completion time measured on a level of one hundredth of a second between Part A and Part B. The smaller the difference, the more efficient participants were in allocating their attention.

An overview of the cognitive tests used in this study is provided in Table 3.6. It summarizes the basal function targeted by the task, the name, source, and overall specification of each test as well as a definition of the resulting measure.

Table 3.6 Overview of cognitive tests in order of presentation

Target function	Test	Operationalisation	Measure
Inhibition	Stroop test (Stroop, 1935; Miyake et al., 2000)	72 trials 60 incongruent & 12 congruent	Difference between reaction time for incongruent vs. congruent
Shifting	Number-letter task (Rogers & Monsell, 1995; Miyake et al., 2000)	32 trials task 1 32 trials task 2 128 trials shifting	Difference between reaction time of repeated vs. shifted
Updating	Keep track (Yntema, 1963; Miyake et al., 2000)	3 trials (4 categories) and 3 trials (5 categories)	Proportion of words recalled correctly
Phonological short-term memory	Forward digit span test (Baddeley, 2003)	14 spans from 3-9 (2 sets per span)	Longest correct span out of all spans
Attention control	Trail making (Tombaugh, 2004)	Part A: connecting consecutive numbers Part B: connecting numbers and letters	Difference between completion time Part B vs. A

3.6.4 Questionnaires and Self-report Data

Two validated questionnaires were included in this study to address the judgement and decision making aspects of Study 2: the General Decision Making Style Inventory (Scott & Bruce, 1995) and the revised Rational-Experiential Inventory (Pacini & Epstein, 1999; 2004). The GDMSI is a 25-item five-point scale differentiating five decision-making styles: *rational*, *intuitive*, *dependent*, *avoidant*, and *spontaneous*. The revised REI-40 is a 40-item five-point scale measuring cognitive style and differentiates two preferred modes of processing information: *rational* and *experiential* (see Section 2.6.3 for a detailed description and Appendix K and Appendix L for the questionnaires and coding).

Additional self-report data on the participants' perception of the rating process were collected via a brief questionnaire in the middle and at the end of each of the two rating sessions (Figure 3.5 and Figure 3.6). Table 3.7 provides an overview of when the raters responded to the self-report items.

Raters were asked to indicate how easy they found it to rate the performances and how confident they were about their ratings on two five-point Likert scales. The response options ranged from *very difficult* to *very easy* and *not at all confident* to *very confident*, respectively. At the end of each rating session, a four-option Likert scale measured how difficult raters found applying the descriptors for each criterion (*very difficult* to *very easy*). Responses had to be justified in one or two sentences (see Figure 3.7).

Table 3.7 Overview of self-report data collection points in experiment

Session 1	Rate 8 performances
	Difficulty and confidence
	Rate 7 performances
	Difficulty and confidence
	Specific difficulty of each criterion and justification
Session 2	Rate 8 performances
	Difficulty and confidence
	Rate 7 performances
	Difficulty and confidence
	Specific difficulty of each criterion and justification

Data collection was completed by a set of exit questions. Raters were asked about their year of birth and length of studies. A final question at the end provided the opportunity to the raters to share a particular observation they have made about the rating process. The specific wording of the question was, "Is there anything you would like to add concerning the rating process as you have experienced it during the course of this

study?” This question was kept open on purpose as the aim was to investigate which features of the rating process were the most noteworthy to the raters. Responses were later used for the case studies.

Figure 3.5 Self-report items after first half of rating session

How did you find rating this first set of performances?

Very difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Very easy
----------------	--------------------	----------------------------	---------------	-----------

How confident about your rating decisions are you?

Not at all confident	Slightly confident	Somewhat confident	Moderately confident	Very confident
----------------------	--------------------	--------------------	----------------------	----------------

You can now take a longer break if you wish or move on to rating the second set of seven performances.

Figure 3.6 Self-report items after second half of rating session

How did you find rating this second set of performances?

Very difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Very easy
----------------	--------------------	----------------------------	---------------	-----------

How confident about your rating decisions are you?

Not at all confident	Slightly confident	Somewhat confident	Moderately confident	Very confident
----------------------	--------------------	--------------------	----------------------	----------------

Figure 3.7 Self-report items and justifications at the end of both rating sessions

Please answer some final questions about the rating criteria.

How did you find each of the criteria to use?

Task achievement

Very difficult	Difficult	Easy	Very easy
----------------	-----------	------	-----------

Why? (1-2 sentences)

Fluency & interaction

Very difficult	Difficult	Easy	Very easy
----------------	-----------	------	-----------

Why? (1-2 sentences)

Lexical & structural range

Very difficult	Difficult	Easy	Very easy
----------------	-----------	------	-----------

Why? (1-2 sentences)

Lexical and structural accuracy

Very difficult	Difficult	Easy	Very easy
----------------	-----------	------	-----------

Why? (1-2 sentences)

3.7 Procedure

3.7.1 Scale Familiarisation Session

All raters (i.e., the novice rater group introduced in Section 3.4.3) completed a two-hour scale familiarisation session prior to rating any performances in a room of the University of Innsbruck. During this training, the basic constructs in the Austrian analytic rating scale were outlined and illustrated with performances. Each session started with raters choosing their pseudonyms for the study, receiving a hard copy of the participant information sheet, and signing the consent form and non-disclosure agreement (see Appendix M and Appendix N). Any questions or concerns about participation in the study were addressed.

A basic knowledge of the principles behind the CEFR and its structure was assumed, as all raters had attended compulsory introductory lectures at the beginning of their studies. The training session began with a brief introduction to the general features of spoken language, and a brief explanation of rater reliability. Then, one criterion at a time was introduced and discussed in the following order: *Fluency and Interaction*, *Accuracy of Spoken Language*, *Range of Spoken Language*, and *Task Achievement*. This sequence was based on my previous experience as a rater trainer and starts with the criteria that are usually easier for the trainees to apply (FLIN, ASL) and then moves on to more challenging criteria (RSL, TA). Following a general introduction for each criterion and a close reading of the scale, raters were shown a performance via a video projector and asked to anonymously rate just the one criterion in question. The ratings were presented and discussed in the plenary without revealing the identity of the raters. Any questions about the nature of the criteria, the wording of descriptors and how they

are generally applied to performances were answered and a benchmark was revealed at the end of each discussion.

The familiarisation session had to be repeated four times, once for each of the cohorts, with participant numbers ranging from two to 16. I paid close attention to replicate the familiarisation sessions by presenting criteria and sample performances in the same order. Even though some of the raters asked questions aimed at behavioural guidelines I withheld any kind of comment that might lead raters to adopt certain strategies. I did so because the documents that are published along the original scales provide no suggestions on how to tackle the rating during a live performance either. The length and format of the familiarisation was also informed by concerns for ecological validity as there has been a growing tendency for in-service teacher trainings in Austria to be no longer than half a day.

Previous research has shown that there can be substantial differences in rating between training sessions and actual rating (e.g., Lumley & McNamara, 1995; Weigle, 1998). To minimise this effect as much as possible, raters would either start rating the first batch of performances on the same day as completing the familiarisation or, at the latest, within three days after the training. In the case of five participants, it was not possible to time the rating sessions so close to the training. They were provided with an extra performance to rate in between the training session and the beginning of the main data collection. All raters, except the five cases just mentioned, completed the rating of all 30 performances within seven days of the familiarisation session.

3.7.2 Rating Sessions

After receiving basic rater training and directly before the first rating session, the raters filled out paper versions of the REI-40 (Pacini & Epstein, 1999) and GDMSI (Scott & Bruce, 1995) in that order. This approach was chosen because the rating of the performances appeared to have a stronger potential to affect the responses on the questionnaires than the other way around. Furthermore, the questionnaires are intended to measure general preferences or tendencies independent of a specific task such as rating. Next, instructions on the rating procedure were presented via Qualtrics and explained in simple language (see Figure 3.8). I highlighted the relevance of the instructions for the quality of the data and emphasised that participants could take breaks at any point during the rating session as long as they had submitted their decisions for the previous performance. Raters received copies of the rating scale, the speaking task and rating forms that they could use to take notes and make comments. After checking that the participants had understood the instructions, they commenced rating.

Each rating session was proctored by me or a research assistant at prearranged times in a computer lab or office at the University. This was deemed important for several reasons. First, this helped ensure that the raters did not use any other applications or resources during the rating and felt more inclined to follow the procedure. Second, it was hoped that proctoring the sessions would motivate raters to concentrate as much as possible. Finally, proctoring the rating sessions was a central concession in the participant consent sheet signed by the parents and speakers who were filmed for this study.

Figure 3.8 Screenshot of instructions prior to rating performances


Instructions

In this set you are going to rate 15 performances. Afterwards you will answer a few short questions.

Please note:

- 1) immediately start playing the videos and never pause them
- 2) enter your rating decisions as soon as you feel you have made them (you can still change your decisions before pressing the >> button, but not after)
- 3) after finishing the rating of a performance and pressing the >> button you can take smaller breaks.

Please enter your pseudonym for this study:



Throughout the rating sessions, JavaScript was running in the background to log mouse events defined as right clicks on rating options in the rating grid. It was, however, not possible to capture the exact moment when raters started viewing the performances as the videos were embedded links to the outside web service YouTube. To improve the quality of the time stamp data, raters were told that their clicks were timestamped and that they had to press the play button of a performance as soon as it had loaded and became visible in the browser. They were also instructed to note down instances when the video did not load immediately or required several clicks to start.

3.7.3 Cognitive Testing

Once raters had completed both rating sessions, an individual session to administer the five cognitive tests (see Section 3.6.3) was scheduled at a time of the raters' choosing. Raters were instructed to select the time slot in such a way that they would

not be too tired from previous courses or pressed for time. Each testing session took between 40 to 50 minutes and was conducted by me.

The five cognitive tests were administered in the following order: 1) Stroop test, 2) Letters-numbers Task, 3) Digit Span test, 4) Keep Track test, and 5) Trail Making. The computer-based tests were administered via a two-monitor setup, where I used one screen to retrieve and start the experiments while the rater sat in front of the other screen that only displayed the tests and stimuli. Each session was concluded with a set of exit questions to collect background data (age, semesters of study). Before closing the session, raters were asked one last open question about their view on the rating process (see also Section 3.6.4 on self-report data).

3.8 Analysis

The data set for this study included both quantitative and qualitative data (see Appendix A). The quantitative data consisted of the analytic scores, awarded by the raters ($N = 39$) on 30 speaking performances, the measures from the cognitive test battery and responses to the questionnaires. Furthermore, observational data from the time stamps and perception of the rating process such as difficulty of rating and confidence as measured by Likert-skale type items were collected and analysed. The qualitative data consisted of raters' justifications regarding the perceived difficulty of using the scale, their handwritten notes in the rating form and their responses to the open-ended question at the end of the cognitive testing session.

The following section of this chapter introduces the measures included in the study and describes the analyses that were carried out.

3.8.1 Rating Data (Study 1)

Following Stemler's recommendations (2004), three types of interrater reliability estimates were analysed: consistency, consensus, and measurement estimates. Following a complete rating plan (Linacre, 1994), each rater ($N = 39$) rated each performance ($N = 30$). As the rating scale included four criteria, a total of 4,680 rating decisions (i.e., 120 decisions per rater) formed the basis of the analysis.

Interrater reliability. Two measures of consensus and consistency were calculated for each rater pairing ($N = 741$). To make this step manageable and reliable, the calculation was based on averaged scores. Thus, all four analytical rating decisions for each performance were averaged and rounded down or up to a full integer value (scores with the first two decimals ending on .49 or below were rounded down, scores whose first two decimals ended on .50 or higher were rounded up).

The percentage of exact agreement and Cohen's kappa (κ) were used as measures of rater consensus. Measures of rater consistency were Pearson's r and Kendall's Tau-b (τ_b). Percentage agreement was calculated via Microsoft EXCEL (2016) and the three other measures (κ , τ_b , r) were calculated via IBM's Statistical Package for Social Sciences (SPSS) for Mac 25. A linear weighting scheme was adopted for the Kappa estimate (Gwet, 2014) to penalize more discrepant rating decisions and reward more adjacent ratings. The freely available SPSS extension Stats Weighted Kappa.spe was used to calculate Cohen's Kappa with linear weights.

Rater severity and consistency. Rating decisions were also examined via an MFRM analysis (Linacre, 1994). All analyses were conducted with the Rasch-based software application FACETS (version 3.80.3, Linacre, n.d.). All performances included in the

experiment were on the same task (“Food waste”, Appendix C.1) and rated with the same analytic rating scale.

The first set of MFRM analyses investigated rater severity and rater fit via a three-facet polytomous rating scale model (RSM, Andrich, 1978). The facets were examinees, raters and criteria:⁶

$$\ln \left[\frac{P_{nijk}}{P_{nij{k-1}}} \right] = \theta_n - \beta_i - \alpha_j - \tau_k$$

where,

- P_{nijk} = probability of examinee n receiving a rating of k from rater j on criterion i ,
- $P_{nij{k-1}}$ = probability of examinee n receiving a rating of $k - 1$ from rater j on criterion i ,
- θ_n = ability of examinee n ,
- β_i = difficulty of the criterion i ,
- α_j = severity of the rater j ,
- τ_k = difficulty of receiving a rating of k relative to $k - 1$ along the rating scale.

The RSM assumes a constant rating scale structure. A score on an 11-category scale is obtained by combining any element of any facet. The candidate facet was floated and positively oriented. Thus, the distribution of this facet was unrestrained and the higher the logit value the higher the candidate score. The other two facets (raters and criteria) were negatively oriented and centred by restraining the logit mean to a value of 0. Raters with positive logit values were more severe and criteria with a higher logit were more difficult.

⁶ Notation for all MFRM analyses adopted from Eckes (2015)

Rater severity is reported as a calibrated Rasch measure of severity (or leniency) in the form of log-odds units (logits). Logit measures in general are interval-level measures and therefore useful for many statistical procedures (Bond & Fox, 2015). Rater severity is an estimate in terms of a latent variable – in this case severity when rating a speaking performance – and therefore associated with some degree of estimation error (Eckes, 2015). The logit rater severity estimate is the distance to the group mean severity which was anchored at 0 of the logit scale.

The extent to which a rater's behaviour fits the expected model is determined via the weighted infit mean-square fit statistic (MS_w) and the unweighted outfit mean-square fit statistic (MS_u). Both measures provide information on how well the raters' observations match the Rasch model expectations (Linacre, 1994; 2020). They are based on squaring the residuals (i.e., the differences between the observed scores and the expected score based on a rater's severity and a candidate's ability) and averaging them across the various elements of a facet. The expected mean-square fit statistic is +1 and values may range from 0 to positive infinity (Linacre, 2020). If values of MS_w or MS_u are above +1, this is an indication of *misfit* suggesting that there is more variation in the observed data than expected by the model. Instances where, for example, a lenient rater provides unexpectedly harsh ratings for a strong candidate can lead to higher mean-square fit statistics. On the other hand, mean-square fit statistics below +1 indicate *overfit* in that less variation than expected occurs in the observed data. Overfit may be caused by central tendency or an avoidance of choosing more extreme bands or failing a candidate (Myford & Wolfe, 2000). Because it is unweighted, MS_u is sensitive to outlying and highly unexpected rating decisions, whereas the weighted MS_w is sensitive to too little model variation (Bond & Fox, 2015).

After a preliminary analysis of rater severity revealed that there were considerable differences between the severity levels of different criteria, the MFRM procedure was run again separately for each criterion. To do so, the model was reduced to two facets (examinees and raters) and run separately by excluding unwanted elements (i.e., criteria) (procedure based on Linacre's recommendations on <https://raschforum.boards.net/>). These analyses yielded five continuous variables – severity for full model, and each of the four criteria TA, FLIN, RSL and ASL.

Rater accuracy. Engelhard (Engelhard, 1996) and Engelhard et al. (2018) suggest the Rater Accuracy Model (RAM) as a further application of MFRM modelling. This approach is based on accuracy scores which are defined as the distance between the reference ratings provided by experts and the score provided by the rater. Accuracy scores constitute a direct measurement of rater accuracy and a dependent variable that can be used to estimate its impact on other facets, for example, criteria and performances (Engelhard, 1996). The advantage of this approach compared to using other measures of accuracy is that it is possible to conceptualize the rater as a facet of the model and, thus, systematically analyse rater differences with regard to a latent accuracy variable (Engelhard, 1996, p. 58). As with the RSM, many facets of the assessment can be included in the RAM and rater variability can be accounted for in terms of the other facets. For instance, certain criteria or performances may just happen to be easier or more difficult to rate accurately than others.

The RAM can be implemented with two basic FACETS models. The observed ratings can be transformed into accuracy scores either via a dichotomous scale (0 = inaccurate, 1 = accurate) or a polytomous scale (scale steps are absolute values of possible differences between observed and reference ratings). Engelhard provides no

recommendations on when to resort to which model except that the polytomous model “may (in some cases) be useful” (1996, p. 69). After a preliminary analysis on basis of the rounded scores and comparing the results, both approaches produced similar outcomes. However, some raters moved several ranks up or down in terms of their accuracy depending on the model. As the main interest for this study was to measure how far the participants approximated the scale use and construct represented by the reference scores, I decided to apply the polytomous model as it appeared to capture similarities in scale use more accurately.

First, the observed scores were recoded following Engelhard’s (1996) instructions. As the maximum absolute difference of a rating decision to the reference score was six bands, exact agreement with the benchmark was recoded to six points. The other ratings were converted into accuracy scores following a partial credit pattern: Missing the reference score by one band was awarded five points, missing it by two bands was awarded four points, missing it by three bands was awarded three points, missing it by four bands was awarded two points and being five bands off the reference score was awarded one point. This conversion also eases the subsequent interpretation of the results as a higher accuracy score becomes indicative of higher rater accuracy. The accuracy scores were then fitted to a three-facet rating scale model (Andrich, 1978) including the facets raters, performance and criteria. The RAM model can be expressed as follows (Engelhard, 1996):

$$\ln \left[\frac{P_{nijk}}{P_{nijk-1}} \right] = \beta_n - \delta_m - \lambda_i - \tau_k$$

where,

- P_{nijk} = probability of rater n assigning an accurate rating to performance m for the criterion i ,
- P_{nijk-1} = probability of rater n assigning an inaccurate rating to performance m for criterion i ,
- β_n = accuracy of rater n ,
- δ_m = difficulty of assigning an accurate rating to performance m ,
- λ_i = difficulty of assigning an accurate rating for criterion i ,
- τ_k = difficulty of accuracy-rating category k relative to category $k-1$.

The results are interpreted in relation to the latent variable rater accuracy. The analysis first investigated parameters of model fit, the accuracy of all raters and difficulty of assigning an accurate rating to the criteria and performances. After this primary analysis, a secondary analysis focused on exploring interaction effects between rater accuracy scores and the other two facets, performances, and criteria. The models for the exploratory bias analyses were:

1. $\ln \left[\frac{P_{nijk}}{P_{nijk-1}} \right] = \beta_n - \delta_m - \lambda_i - \beta_n \lambda_i - \tau_k$
where $\beta_n \lambda_i$ represents the rater-by-criterion interaction parameter or bias term,
2. $\ln \left[\frac{P_{nijk}}{P_{nijk-1}} \right] = \beta_n - \delta_m - \lambda_i - \beta_n \delta_m - \tau_k$
where $\beta_n \delta_m$ represents the rater-by-performance interaction parameter, and
3. $\ln \left[\frac{P_{nijk}}{P_{nijk-1}} \right] = \beta_n - \delta_m - \lambda_i - \beta_n \delta_m \lambda_i - \tau_k$
where $\beta_n \delta_m \lambda_i$ represents the rater-by-performance-by-criterion interaction parameter.

Finally, after the first RAM analysis and bias analysis, four separate criterion-specific analyses were carried out to capture some of the finer differences in terms of rater accuracy. To do so, the three-facet model was reduced to two facets (raters and performances) and only included data for the selected element (i.e., criterion) in the estimation. This approach yielded a total of five continuous variables (accuracy for the full model and for each criterion TA, FLIN, RSL and ASL).

3.8.2 Thematic Analysis (Study 1)

Participants were prompted to self-report on several dimensions at multiple points of the experiment (see also 3.6.4). The items focussed on how difficult raters found it to rate in general, how confident they were about their decisions and, at the end of each session, how difficult they found rating each criterion. The responses to the Likert-type items were exported from the online rating platform and analysed as quantitative data. The 312 statements justifying why a certain criterion was either difficult or easy to apply (4 justifications x 2 rating sessions x 39 raters) were transferred to the qualitative data analysis software MAXQDA Plus 2018 (Verbi Software).

The statements were analysed following an exploratory thematic analysis (Braun & Clarke, 2006). Thematic analysis is a commonly used approach for “identifying, analysing, and reporting patterns (themes) within data” (p. 79) and increasingly has come to be viewed as a method in its own right (Joffe, 2012). Viewing the raters’ responses to the open-ended item as a “proxy of experience” (Bernard & Ryan, 1998), the analysis was content-driven in that no predetermined theory was superimposed on the responses prior to the analysis (Guest et al., 2012). Instead, themes were regarded as emergent and developed by frequent rereading, coding, and rearranging of the segments, until a rich overall description of the entire data set was reached.

The analysis mostly followed the phases as suggested in Braun and Clarke (2006). First, a first layer of codes was added to each statement to capture information on the criterion and its perceived difficulty (data from Likert-type item). Next, all 312 segments were read with a view to identifying underlying themes. Some of the statements included more than one theme and were split into several segments, leading

to a total of 376 segments. After reading the segments several times, a rough coding scheme was devised.

Once a rough scheme had been established, the Likert-scale responses were used to pile segments according to whether they were part of justifications of why a criterion was ‘(very) easy’ or ‘(very) difficult’ to rate. The code book was then refined by rereading the newly grouped segments and combining or differentiating codes whenever necessary. After collating segments by perceived difficulty, segments were revisited grouped by code to review each code. The final code book consisted of two major code families: facilitators ($n = 134$) and inhibitors of rating ($n = 174$) (Appendix I).

3.8.3 Time Stamp Data (Study 2)

Throughout the two rating sessions delivered via Qualtrics, JavaScript recorded timestamps for each mouse event. Three measures were extrapolated from the timestamp data: (1) deliberation time (DT), (2) time to first decision (TTFD), and (3) revision count (RC). DT was operationalized as the length of time spent with a performance minus the length of the performance. TTFD was defined as the time interval between transitioning to rate the next performance and the time stamp for the first mouse click on a rating band. RC indicated how many times a rater revised rating decisions throughout the rating sessions.

Several steps were taken to optimize the quality of the time stamp data. First, raters were instructed to revise their decisions as many times as they wished before submitting. Raters were also asked to record noticeable delays with video buffering in their rating forms. This was to be able to account for irregularities in the data due to

fluctuations in buffering speed. Finally, participants were instructed to not take breaks before submitting their rating decisions but to instead immediately submit their ratings when they had finalized them and take a break while on a transition page in the online questionnaire environment.

After exporting the time stamps into EXCEL, the quality of the data was inspected for unexpected or extreme observations. In the following section, the quality of the data and definition of thresholds for useful observations will be described in more detail.

Deliberation time (DT). Deliberation time was defined as the time a rater spent on rating a performance. Variations in performance lengths were controlled for by subtracting the length of each performance from the time a rater spent on rating the performance (i.e., timestamp of starting to view a performance to timestamp of submitting all four rating decisions). Negative DT values indicated that a rater submitted the rating before the performance had completed, while positive DT values indicated that raters continued to think about their rating decisions even after the performance had ended.

As the mean length of all performances was about five minutes, the lower threshold for useful observations was set at two and a half minutes (150 seconds). This would identify those instances where raters failed to at least watch about half of a performance before submitting their ratings and moving on to the next performance. Overall, there were nine instances below the 150-second threshold: four in Session 1 and five in Session 2. Three observations were produced by the same rater, Helen. All nine extremely short DTs were excluded from the analysis of rating behaviours and defined as missing in the SPSS file.

Extremely long DTs were defined as longer than ten minutes or double the average length of a performance. Seven raters needed longer than ten minutes for at least one performance in the first half of Session 1. Longer DT, however, were rare in Session 2 ($n = 3$). One rater, Lucy, needed ten minutes for ten out of 15 performances in the first rating session. As taking extremely long for a rating decision cannot as such be considered problematic, only one extreme observation (1,013.32 s) from participant Paige was removed as an outlier from the data set. In light of her mean overall deliberation time ($M = 154.37$) it was clear that this rater took quite long for her rating decisions, but in this particular case, she might have forgot to move on to the intermission screen before taking a break.

Time to first decision (TTFD). This variable was defined as time duration between starting to rate a performance and entering the first decision by selecting a box in the online rating grid representing the rating scale. When inspecting the exported time stamps and deliberation times, seven data points were identified as outliers or missing. Some of the extreme outliers in terms of DT (see previous section) also displayed irregularities in terms of TTFD and it did not seem reasonable to include the TTFD time stamp if the rater overall did not spend an adequate amount of time with a performance. After removing the seven extreme observations, all other data points were exported into SPSS.

Revision count (RC). Revisions were identified by looking at the number of clicks on the scale bands. To calculate the number of revisions, the minimum number of clicks required for the rating session (i.e., number of performances x four criteria) was subtracted from the total number of mouse clicks recorded for that session. A value of 0 indicated no revisions. The minimum number of timestamps required for all 39 raters

to perform the rating of 30 performances with four criteria was 4,680. Overall, JavaScript recorded 5,712 timestamps, implying that 1,032 revisions were made during the rating sessions. However, four timestamps turned out as missing which affected the rating by three raters on one performance. These were instances where decisions had been taken extremely fast and they had also been excluded from the analysis of the other two rating behaviour metrics (DT, TTFD). When calculating the average revisions per criterion and rater, those missing four counts were added. This was the least intrusive measure to neutralise the missing data and it is unlikely that raters had the intention of revising a decision in these instances as they seemed to have decided to just quickly get the performance out of the way or might have made a mistake when entering the data. Time stamp data was exported into SPSS.

3.8.4 Cognitive Test Scores (Study 2)

The purpose of collecting and analysing the cognitive measures was to investigate whether components of executive function (inhibition, shifting, updating), phonological short-term memory and attention displayed a relationship with the dependent variables of rating quality (i.e., rater severity, consistency, and accuracy) or rating behaviour (i.e., time to first decision, deliberation time and number of revisions). The data was collected during individual sessions with each rater. After importing the scores into SPSS, the three variables, which measured performance in terms of time (Stroop, Letters-numbers and Trail Making) were reverse coded so that a higher score represented a stronger performance rather than a slower and weaker performance. This was achieved by subtracting the actual time a participant needed from the maximum value obtained in the sample. As a result, the direction of all cognitive variables was aligned (i.e., higher score means stronger performance). The

Keep Track data which was a count of recalled words out of 27 was transformed into a proportion.

3.8.5 Questionnaire Data (Study 2)

Both questionnaires were completed on paper (see Appendix K.1 and Appendix L.1). The data were transferred into EXCEL and imported into SPSS. Despite the clear layout of both questionnaires, there were four missing responses for the REI-40 (.026%) and three for the GDMSI (.031%). These were defined as missing data in the SPSS file.

GDMSI. Calculating the DMS profiles for each participant is straightforward as no items require reverse coding. The five types of DMS are represented by five items each in the 25-item questionnaire. Following the item map (Appendix K.2), mean scores were calculated for DMS, resulting in five DMS mean scores (one per style) per participant.

REI-40. Negatively oriented items in the REI-40 were reflected in SPSS (see Appendix L.2). The REI-40 comprises two main scales (rationality and experientiality with 20 items each), which can further be split into four ten-item subscales (rational ability and engagement, experiential ability, and engagement). Mean scores were calculated for the two main scales and the four subscales, leaving six variables for initial analysis.

The Processing Style Influence (PSI) score (Gunnell & Ceci, 2010) required establishing total scores for the two main scales, rationality and experientiality. Next, the PSI for each participant was calculated following this formula:

$$\text{PSI} = [(Mdn \text{ cohort rationality}) - (\text{individual rationality score})] + [(\text{individual experientiality score}) - (Mdn \text{ cohort experientiality})]$$

The PSI score assumes that one preferred mode of processing overrules the other. PSI scores can be positive, which indicates that a person is primarily experientially influenced in their processing, or negative, which identifies persons with a preference for rational processing (Gunnell & Ceci, 2010).

3.8.6 Inspection and Transformation of Variables (Study 2)

Prior to conducting the multiple correlational analyses for Study 2, all continuous variables (see Appendix A for Variable Map) were inspected. Features of an approximate normal distribution and outliers were investigated for each variable by looking at the normal Q-Q plots, histograms and box plots, standardised values of skewness and kurtosis as well as the Shapiro-Wilks statistic. As a normal distribution could not be established for several variables (e.g., Digit Span, Trail Making, Spontaneous DMS, Avoidant DMS) and as sample size was relatively small, all correlations were calculated using the non-parametric Spearman rank correlation.

The correlational analyses, which form the main body of Study 2, were supplemented with bootstrap confidence intervals to indicate the accuracy and stability of the correlation coefficients (LaFlair et al., 2015). 95% confidence intervals were calculated via a bias-corrected-and-accelerated bootstrap method based on 2,000 random resamples.

The stepwise Holm-Bonferroni procedure (Holm, 1979; Aickin & Gensler, 1996) was adopted to manage the Type I family-wise error rate for the multiple hypotheses tests. This procedure improves statistical power in that it is less conservative than the original Bonferroni method, yet conservative enough to avoid Type I errors of falsely

rejecting the null hypothesis. The corrected p -values were obtained through an EXCEL calculator (Gaetano, 2018), and corrections were calculated for 12 comparisons per dependent variable (i.e., cognitive test variables, questionnaire variables). Intercorrelations between dependent variables or questionnaire subscales remained uncorrected.

3.8.7 Case studies (Study 3)

A multiple-case study approach (Stake, 2006) was chosen to investigate several substantially interesting raters from Study 1 and 2. Case studies can be viewed as a strategy of investigating a phenomenon (Casanave, 2015) with a view to “illustrating [it] in very vivid, detailed, and highly contextualized ways from different perspectives” (Duff, 2019, p. 145) and “enhance our understanding” (Casanave, 2015, p. 120). The goal of the case studies in this study was to address the shortcoming of predominantly quantitative rater cognition research (see literature review in Chapter 2), and, through an *instrumental* approach of selecting and investigating cases (Stake 2005, p. 445), learn more about the rating process.

A central step in case study research is to decide on a rationale for case selection (Casanave, 2015). The data collected for this project had the potential to be used to support various approaches to selecting interesting raters. I decided to take a maximum variation, or extreme case stance and juxtapose extremely successful and unsuccessful raters to demonstrate the diversity of the sample along the multiple sources of data included in the research (see also Jahnukainen, 2010). In the context of a standardized examination, it can be argued that rater accuracy (i.e., agreement with standards) is of greater importance than rater severity. Thus, the main criterion for selecting cases was the participants’ rating accuracy – overall and per criterion – as determined through

MFRM analysis based on the Rater Accuracy Model (Engelhard, 1996; see also Section 3.8.1). In a second step, the quality of the handwritten notes taken during the rating and rating behaviour metrics were taken into consideration. Finally, aspects such as rater fit, decision-making profiles or preferences were also considered to have some variability in the in-depth case descriptions.

After selecting the four cases, data for each individual case were amalgamated for a first within-case analysis. This meant breaking group-level statistics down to individual-level ranks and comparisons. Furthermore, cases were supplemented with additional observational data gathered throughout the experiment to develop a rich description in accordance with the case study approach. Supplementary sources of data were the participants' responses to the open-ended exit question asked at the end of the data collection and the handwritten notes that the raters took during the rating sessions, which provided insights into the features raters noticed. In a last step, all four cases were compared to see how they varied, and if there were any commonalities to be found among them.

3.9 Summary

The chapter set out to outline the methodological approach and procedures employed in this research. It provided a rationale for employing a mixed-methods stance towards data collection and analysis as well as an overview of the research design which consists of three consecutive studies. The remaining part of the chapter was dedicated to describing the research context, participants, instruments, material development, procedures, and data analysis for each of the three studies. The following three chapters will now present the results.

4 Results I: Rating Quality and Scale Use

4.1 Outline

This chapter presents the results of the analyses regarding the scale use of a group of novice raters. Via measures based on classical test theory (CTT) and scaled estimates (MFRM), it will investigate the warrant that raters rate reliably linked to the evaluation inference within the argument-based validation framework (Knoch & Chapelle, 2017). Thus, overall rating quality will be established through investigating rater consistency, consensus, fit and bias. In line with Knoch and Chapelle's suggestions (2017, p. 7), the analysis will also include data on how comfortable raters were with applying the descriptors from each scale dimension and how confident they were in their decisions. In addition, the analysis will also present estimates of overall rater accuracy and criterion-specific accuracy based on the Rater Accuracy Model (e.g., Engelhard, 1996). These will be presented in more detail as the extent to which Austrian teachers, or in this case novice raters, would agree with the exit-level standards as illustrated by the reference scores could be considered more important than inter-rater reliability.

Section 4.2 presents findings on the consistency and consensus among the overall scores provided by the raters based on the CTT approach. Section 4.3 reports on the results of the first MFRM analysis (rater severity model, RSM) and focuses on rater severity, rater fit and model fit, as well as subsequent analyses for each criterion. Next, Section 4.4 presents results of an investigation of rater accuracy by first comparing the scores produced by the raters with reference scores in terms of consistency and consensus. Section 4.5 summarizes results from an MFRM analysis based on the Rater

Accuracy Model. Finally, Section 4.6 will report on the results of self-report data regarding rater perceptions.

4.2 Interrater Reliability (CTT Approach)

There were 741 possible pairings for an analysis of consistency and consensus of this rater group ($N = 39$). For a first look at the data, three measures of interrater reliability were calculated on basis of the rounded scores: 1) percentage exact agreement, 2) weighted Cohen's kappa (κ), and 3) Kendall's Tau-b (τ_b). Table 4.1 provides sample data of the ten most extreme pairings ranked by kappa, with the five strongest pairs on the top and the five weakest pairs towards the bottom. This illustrates the fundamental differences between measures of consensus, such as the percentage exact agreement and kappa, and measures of consistency, in this case Tau-b.

Several raters like Daisy, Amy and Penelope appear in pairs at both ends of the spectrum. Daisy and Amy, for instance, obtained the second highest kappa value ($\kappa = .628, p < .001$) of all possible pairings and agreed exactly in nearly two thirds of the ratings (63.33%). In terms of consistency, there was also a considerable correspondence in how these raters ranked the performances ($\tau_b = .782, p < .001$). This pair, thus, achieved considerable consensus and quite strong consistency in their ratings.

When Daisy and Amber's ratings are compared, on the other hand, the indices of consensus and consistency are quite different from each other. Even though this analysis was based on rounded scores, which could be expected to even out some discrepancies, the pair only agreed in the case of one out of 30 speakers (3.33%, $\kappa = .084, p = .04$). Interestingly, their consistency metric is notably quite strong ($\tau_b = .634,$

$p < .001$). While this pair hardly agreed in terms of average overall score, their consistency is not too far removed from .70 which is commonly considered a lower-bound threshold level for acceptable interrater consistency (Salkind, 2010). Comparisons such as this highlight the need to rely on several measures when investigating interrater reliability as using just one measure can easily lead to a distorted interpretation of the data.

Table 4.1 Comparison of most extreme rater pairs in terms of consensus and consistency

Rater 1	Rater 2	<i>N</i> exact agreement	% exact agreement	κ	τ_b
Eliza	Phoebe	20	66.67%	.67	.72
Amy	Daisy	19	63.33%	.63	.78
Lucy	Susan	16	53.33%	.62	.67
Esme	Kimberly	17	56.67%	.60	.66
Dorothy	Penelope	13	43.33%	.60	.71
...					
Daisy	Helen	3	10.00%	.09°	.57
Amber	Daisy	1	3.33%	.08°	.63
Amy	Betty	1	3.33%	.08°	.48
Betty	Daisy	2	6.67%	.08°	.50
Helen	Penelope	6	20.00%	.05°	.09°

Note. All measures except those marked by ° are significant at the .01 level. Data is based on rounded scores. κ = Cohen's kappa, τ_b = Kendall's tau-b.

The descriptive statistics indicate that there is a great variability across all indices of inter-rater reliability between the 741 possible pairs of raters (see Table 4.2). It is noteworthy that some pairings attain extremely low minimum values (exact = 3%, κ = .050, τ_b = .09); 20 percent of all pairs remain below 21% exact agreement and 13 percent had a kappa below .21. Only around 15 percent (for exact agreement) and 26

percent (for kappa) are above values of 41% or .41, respectively. As a result, the mean indices of consensus are not particularly strong ($M_k = .33$, $SD = .10$; $M_{\% \text{ agreement}} = 30.49$, $SD = 10.46$). As far as consistency is concerned, this cohort of raters achieves a somewhat reasonable overall mean ($M_{\text{tau-b}} = .49$, $SD = .10$). Most pairs obtained a value of tau-b above .41 (62.1 %) or above .61 (17.3 %). Thus, the range of variation within the 741 pairings is great, indicating a heterogeneous approach towards rating the performances. Moreover, raters tended to agree more in how they ranked the performances (i.e., their consistency) rather than in the actual rating decisions (i.e., consensus), with about a fifth of raters approaching acceptable levels of consistency.

Table 4.2 Descriptive statistics of inter-rater reliability measures

Statistic	Exact agreement (%)	κ	τ_b
M (SE)	30.49 (0.38)	.33 (.00)	.50 (.00)
Mdn	30	.33	.5
SD	10.46	.11	.11
Min	3	.05	.09
Max	67	.67	.78

Note. ^a Multiple modes exist. $N = 741$ pairs. The smallest value is shown. Data based on rounded scores. κ = Cohen's kappa, τ_b = Kendall's tau-b.

4.3 Rater Severity and Consistency

4.3.1 Wright Map

Rater severity and consistency was estimated by a three-facet MFRM analysis based on the Rating Scale Model (RSM, see Methodology Section 3.8.1). The Wright Map (in

Figure 4.1) shows the relative calibration of the elements in each facet. Elements in each facet (examinee, raters, criterion), which are displayed horizontally, are mapped onto a vertical scale and extreme values within each facet are either towards the upper or lower end of the map. The first column shows the logit scale onto which the facets have been calibrated. The examinee facet was defined as floating, while the other two facets are centred at a mean value of 0. Floating the examinee facet and centring other facets is the standard procedure for a first general analysis of a set of scores from a typical performance assessment (Eckes, 2015; Green, 2013). The examinee facet is positively oriented, whereas the two other facets are negatively oriented. Consequently, the candidates located at the upper end of the scale are also the candidates that received the highest scores. Conversely, the other two facets (raters, criterion) must be interpreted the other way round; the higher up a data point is located in the Wright map for those two columns, the more severe the rater was in their observations or the more difficult it was to receive a higher score for that criterion.

Examinee abilities ranged from -0.34 to 3.95 logits. A mean fair average score of 7.45 logits indicates that the examinee cohort was relatively strong overall. However, these findings are in line with the results from the previous rating round with expert raters and was to be expected since many of the speakers recruited for the simulated speaking assessments were in their final year of upper-secondary school or at university.

As can be seen from the third column, raters, there is considerable variability in terms of rater severity. With logit values ranging from -0.90 to 0.82 the difference between raters spans almost 2 logits (1.72), covering a little less than half (40.1%) of the range observed in the examinee facet. The fourth column for the facet criterion is a first

indication that the criteria were not of equal difficulty; the two language related criteria RSL and ASL appear to be rated more severely than the criteria FLIN and TA.

Figure 4.1 Measurement rulers for RSM analysis

Measr	+examinee	-raters	-criterion	Scale
4	+	+	+	+(10)
	*			
3	+	+	+	+
	*			
	*			
	*			
2	+	+	+	+
	**			
	*			
	*			
	*			
	**			
	**			

1	+	+	+	+
	**			
	*	*		
	*	**		
	*	**		
	*	**		
	****	****		
	*	*		
	*	***	LSA	
	*	*	LSR	
0	*	*****		*
	*	**	FLIN TA	6
	*	***		
	*	***		
	*	***		
	*	*		
	*	*		
	*	**		
-1	+	+	+	+(2)
Measr	* = 1	* = 1	-criterion	Scale

4.3.2 Group-level Statistics

Table 4.3 summarizes a range of group-level statistics as produced in the facet measurement reports of FACETS. The upper half of the table presents descriptive statistics for the estimates of each facet (M , SD , M of SE , root mean-square error $RMSE$, and adjusted SD) while the other half lists separation statistics which is useful to investigate the variability of the data for each facet (homogeneity index, separation ratio, separation index and separation reliability). The descriptive statistics and adjusted standard deviations indicate that the estimates are most varied for the

examinee facet ($SD = 0.90$) and much less so for the other two facets, raters ($SD = 0.43$) and criterion ($SD = 0.12$). The average measurement of the location of each element on the latent variable, i.e., the precision of the estimates, as indicated by the standard error (SE) is quite precise (Linacre, 2020).

Table 4.3 Summary statistic for MFRM analysis of rater severity and fit

Statistic	Examinee	Raters	Criterion
<i>M</i> (measure)	1.17	.00 ^a	.00 ^a
<i>SD</i> (measure)	0.90	0.43	0.12
<i>M</i> (SE)	0.08	0.08	0.03
RMSE	0.08	0.08	0.03
Adj. (true) <i>SD</i>	0.89	0.43	0.12
Homogeneity index (chi-square)	2,998.8*	990*	80.4*
df	29	38	3
Separation ratio	11.44	5.01	4.37
Separation (strata) index	15.58	7.01	6.16
Reliability of separation	0.99	0.96	0.95

Note. RMSE = root mean-square measurement error. ^a The benchmark and criterion facets were centered and constrained to have a mean element measure of zero. * $p < .05$

The separation indices need to be interpreted in turn for each facet as their interpretation is contingent on the facet itself. Looking first at the examinee facet, there is evidence that the speakers in the sample were of varying ability levels. The homogeneity index, which tests for statistically significant differences in logits between elements, confirms that at least two speakers differ significantly in their ability estimates ($\chi^2 = 2,998.8$, $df = 29$, $p < .005$). The separation ratio, which relates examinee ability measures to the precision of measurement can take on a value between 0 and infinity (Eckes, 2015). The higher this index the greater the variation between the elements of a facet. Estimates of speaker ability were about 11 times larger than the precision of these estimates. According to the separation index, examinees could be separated out into 15 statistically distinct levels of ability, which

is slightly more than the ten bands available in the rating scale that was used. According to Eckes (2015, p. 65), one explanation for this quite high separation might be the large “true” standard deviation of 0.89. Finally, the reliability of separation reaches the theoretical maximum of .99. All indices point towards some variability in ability levels. Following Myford and Wolfe (2004), it is also unlikely based on these metrics that there is a group-level effect of central tendency or clustering of observations around the middle bands of the rating scale.

While larger separation indices are somewhat desirable for the examinee facet as this shows that the measurement procedure successfully discriminated stronger and weaker candidates, these indices need to be interpreted somewhat differently for the other two facets. For raters, the separation ratio (5.01) and the separation strata (7.01) suggest that there are considerable severity differences between the raters (i.e., five times greater than the error of measurement) and that there are at least seven distinct levels of severity within this group. The separation reliability of .96 further confirms that the difference in rater severity is reliable and that there is a high degree of dissimilarity in the scoring decisions of the raters. All indices of the rater facet imply a degree of heterogeneity in rater decision making in this cohort.

Finally, the separation ratio (4.37) and separation index (6.16) of the criterion facet are quite close to the number of elements in the facet. There appear to be six statistically distinct levels of difficulty which is less than the ten bands available in the scale. This is to be expected as the analysis leading up to creating the fair score benchmarks has already shown that there was a distinct lack of weak performances in the set of speakers used for this study.

4.3.3 Rater Facet Measurement Results

After investigating the group-level statistical indicators, the focus shifts to the rater facet. Table 4.4 presents the logit locations of the raters in terms of their severity and the infit and outfit mean-square values. Severity estimates range from 0.82 logits ($SE = .08$) for Amber, who was the most severe rater with a mean score of 6.47 on the ten-step scale to -0.90 ($SE = .09$) for the most lenient rater Daisy ($M = 8.45$). The relatively small standard errors (SE between .08 and .09) reported in the fourth column indicate that the calibration of each rater's location can be considered quite precise (Eckes, 2015; Engelhard & Wind, 2018). The minimum and maximum values of infit and outfit mean-square fit statistics (MS_w and MS_u) columns range from 0.65 and 0.64 to 1.85 and 1.87, respectively. Finally, two columns (t_w and t_u) list the transformed infit and outfit statistics as t -statistic. The standardized fit statistic can be used as a significance test for identifying raters whose infit and outfit is statistically significant (t smaller than -2.0, or t larger than 2.0) at the 0.05 level of significance.

Table 4.4 Measurement report for rater facet

Rater	M	Severity (Logits)	SE	MS_w	t_w	MS_u	t_u
Amber	6.47	0.82	0.08	0.94	-0.41	0.93	-0.48
Penelope	6.64	0.68	0.08	1.56	3.71	1.81	5.05
Betty	6.65	0.67	0.08	0.77	-1.91	0.76	-2.00
Nancy	6.76	0.58	0.08	0.77	-1.90	0.76	-1.95
Helen	6.78	0.56	0.08	0.77	-1.93	0.78	-1.84
Dorothy	6.85	0.51	0.08	0.75	-2.05	0.74	-2.12
Paige	6.92	0.45	0.08	1.14	1.07	1.15	1.13
Mary	6.98	0.40	0.08	1.36	2.56	1.34	2.41
Maisie	6.97	0.40	0.08	0.81	-1.51	0.81	-1.55
Maddison	7.01	0.38	0.08	1.85	5.40	1.87	5.41
Zara	7.03	0.35	0.08	1.18	1.36	1.16	1.24
North	7.12	0.28	0.08	0.85	-1.22	1.04	0.37
Hendrix	7.16	0.25	0.08	0.86	-1.08	0.88	-0.93

Tyler	7.17	0.24	0.08	1.24	1.79	1.32	2.31
Willow	7.23	0.19	0.08	0.91	-0.67	0.95	-0.34
Holly	7.31	0.12	0.08	1.47	3.30	1.50	3.42
Chloe	7.42	0.03	0.08	1.23	1.72	1.24	1.77
Donny	7.42	0.03	0.08	0.91	-0.67	0.93	-0.48
Susan	7.43	0.02	0.08	0.67	-2.92	0.66	-3.00
Zachary	7.46	0.00	0.08	1.01	0.12	1.01	0.15
Phoebe	7.47	-0.02	0.08	0.87	-1.07	0.87	-1.02
Lucy	7.51	-0.05	0.08	0.65	-3.15	0.75	-2.07
Eliza	7.58	-0.11	0.08	0.88	-0.99	0.85	-1.17
Margarete	7.62	-0.15	0.08	0.87	-1.08	0.90	-0.79
Violet	7.64	-0.16	0.08	0.97	-0.22	1.01	0.14
Kimberly	7.66	-0.17	0.08	1.03	0.26	0.99	-0.05
Rosie	7.69	-0.20	0.08	1.63	4.26	1.66	4.28
Layla	7.77	-0.27	0.09	0.79	-1.75	0.82	-1.47
Scarlett	7.81	-0.31	0.09	1.08	0.67	1.06	0.52
Stormi	7.82	-0.32	0.09	0.68	-2.87	0.72	-2.35
Jennifer	7.87	-0.36	0.09	0.65	-3.11	0.64	-3.14
Esme	7.88	-0.37	0.09	1.33	2.39	1.45	3.07
Sally	7.89	-0.38	0.09	1.30	2.23	1.30	2.12
Lexi	7.98	-0.46	0.09	0.70	-2.67	0.88	-0.91
Alice	8.10	-0.57	0.09	0.92	-0.59	0.97	-0.15
Deborah	8.20	-0.66	0.09	0.96	-0.30	1.16	1.20
Megan	8.20	-0.66	0.09	0.90	-0.75	0.85	-1.09
Amy	8.43	-0.87	0.09	0.72	-2.35	0.67	-2.61
Daisy	8.45	-0.90	0.09	0.79	-1.75	0.77	-1.69
<i>M</i>	7.44	0.00	0.08	0.99	-0.21	1.02	0.04
<i>SD</i>	0.51	0.44	0.00	0.29	2.16	0.31	2.19

Note. MS_w = mean-square infit statistic. t_w = standardized infit statistic. MS_u = mean-square outfit statistic. t_u = standardized outfit statistic. Raters are ordered by severity measure from high (more severe) to low (more lenient).

Various suggestions for upper and lower threshold limits for mean-square fit statistics are suggested in the MFRM literature. While considering mean-square values between 0.5 and 1.5 as “productive for measurement” (p. 272), Linacre (2020) recommends an upper limit of 1.2 and lower limit of 0.4 for scores based on judgement. Narrower thresholds that are used quite commonly have also been suggested and range from

0.70 to 1.30 (Bond & Fox, 2015; Eckes, 2009; McNamara, 1996). Eckes (2015) as well as Bond and Fox (2015) highlight the need to critically interpret and adopt upper and lower thresholds depending on the test context or type of data used in the Rasch model. There is also evidence that fit statistics are sensitive to sample size and that they may have to be adjusted in very large samples (Wu & Adams, 2007). Table 4.5 below summarizes the number and percentages of raters in terms of their fit indices. While about 70% of the participants could be fit to the model within a narrowly defined threshold (0.7 to 1.3 logits), there are several raters that display significant overfit (for MS_w 5 or 12.8%; for MS_u 3 or 7.7%) and an even larger group with significant misfit (for MS_w 7 or 17.9%; for MS_u 8 or 20.5%).

Table 4.5 Summary of raters (number and percent) within and outside of fit thresholds

Fit thresholds	MS_w		MS_u	
	<i>N</i>	%	<i>N</i>	%
fit < 0.7, $t > \pm 2.0$	5	12.8	3	7.7
0.7 < fit < 1.3	27	69.2	28	71.8
fit > 1.3, $t > \pm 2.0$	7	17.9	8	20.5
Total	39	100.0	39	100.0

4.3.4 Criterion Measurement Results

Next, the results for the four elements of the criterion facet (TA Task Achievement, FLIN Fluency and Interaction, RSL Range of Spoken Language and ASL Accuracy of Spoken Language) will be presented in more detail (see Table 4.6). As the Wright map already indicated, a candidate was more likely to attain a higher score for the two criteria FLIN and TA than for ASL or RSL. The language-related criteria ASL and RSL were estimated to be more difficult with severity measures of 0.17 ($SE = .03$) and 0.06 ($SE = .03$), respectively, than FLIN (logit -0.08, $SE = .03$) and TA, (logit -0.14, $SE = .03$). The precision of the model as indicated by the standard error is very high

which is likely due to the large number of responses ($N = 1,170$). In order to investigate the null hypothesis of equal difficulty, the Wald statistic as described in Eckes (2016, p. 61) confirmed that only the criteria TA and FLIN did not differ significantly in terms of relative difficulty $t_{TA,FLIN}(2338) = 1.41$, ns. All other comparisons between the relative difficulty of the criteria yielded significant results ($t_{ASL,RSL}(2338) = 2.59$, $p < 0.05$ and $t_{ASL,FLIN}(2338) = 3.30$, $p < 0.01$).

The fit statistics suggest issues with measurement accuracy. The mean-square fit statistics indicate significant misfit for the criterion TA ($MS_w TA = 1.32$, $MS_u TA = 1.48$), which is further substantiated with a positive t -statistic of 7.07. The criteria RSL and FLIN also both display a significant departure from the model expectations with fit indices clearly below 1 ($MS_w RSL = 0.83$, $MS_u RSL = 0.82$; $MS_w FLIN = 0.86$, $MS_u FLIN = 0.84$) and t -statistics below -2 ($t_w RSL = -4.28$, $t_u RSL = -4.58$; $t_w FLIN = -3.60$, $t_u FLIN = -4.06$). However, while the data for these two elements do not fit the model perfectly, they are still useful (Linacre, 2003). Linacre (2020) suggests various explanations for misfit and overfit of rating scale categories. Misfit, for instance, may be a result of extreme category overuse while overfit might be due to the overuse of middle categories and central tendencies. Problematic fit statistics also raise questions about the unidimensionality of the construct in relation to the scale (Eckes, 2015; McNamara, 1996). At least at the group level, there is little evidence for a central tendency and overuse of the middle categories. If that were the case, the values of the separation statistics and reliability would have been lower in the examinee facet (see Table 4.3 above). Thus, the reason for the misfit might be due to a differential functioning of the criteria.

Table 4.6 Criterion measurement report

Criterion	Obs. Score	<i>M</i>	Fair <i>M</i>	Measure	<i>SE</i>	<i>MS_w</i>	<i>t_w</i>	<i>MS_u</i>	<i>t_u</i>
ASL	8481	7.25	7.29	0.17	.03	0.98	-0.49	0.96	-0.94
RSL	8635	7.38	7.44	0.06	.03	0.83	-4.28	0.82	-4.58
FLIN	8821	7.54	7.62	-0.08	.03	0.86	-3.60	0.84	-4.06
TA	8907	7.61	7.71	-0.14	.03	1.32	7.07	1.48	9.00

Note. *MS_w* = mean-square infit statistic. *t_w* = standardized infit statistic. *MS_u* = mean-square outfit statistic. *t_u* = standardized outfit statistic. *TA* = Task achievement. *FLIN* = Fluency and interaction. *RSL* = Range of spoken language. *ASL* = Accuracy of spoken language. Estimates based on 1,170 observations per criterion.

4.3.5 Global Model Fit

A final source of information on Rasch model fit are the residuals produced during the main FACETS analysis. While the software repeats the iterations to fit the scoring data to the Rasch model, instances where raters provided an unexpected score are highlighted by the software. Table 4.7 provides an overview of all unexpected observations based on the RSM model with an absolute residual value of 3.0 and above ($N = 34$). Seventeen of the 39 participants did not produce any unexpected ratings and there were several raters like Amber for whom only one out of 120 rating decisions produced a high residual. However, there were also three raters (Maddison, Penelope, and Rosie) who each provided three highly unexpected rating decisions. Moreover, there were some speakers (e.g., 40, 27, 22, 37) whose performances seem to be linked to unexpected ratings and might have been more difficult to rate consistently than other performances. Unsurprisingly in light of the measurement results for the criterion facet (Table 4.6), ratings for TA were far more prone to elicit unexpected ratings than any of the other criteria. There were 27 (79.4%) unexpected rating decisions associated with TA, compared to four for ASL (14.8%), two for FLIN (5.9%) and just one for RSL (2.9%). According to Linacre (2020), such a

concentration of residuals around certain facet elements – in this case the criterion TA – are a sign of local misfit and deviation of model expectations. However, while they are likely to negatively affect the overall global model fit, the total number of residuals ≥ 3.0 in this set is still below the 1% maximum threshold suggested by Linacre ($n = 34$, 0.72% of 4,680 observations). While the analysis revealed issues with certain elements, the global model fit can still be considered satisfactory.

Table 4.7 Residuals from FACETS analysis (listed by rater)

Raters	Obs. Score	Exp. Score	Residual (Obs.-exp.)	Standard. Residual	Candidate	Criterion
Amber	5	8.6	-3.6	-3.5	40	TA
Chloe	4	8.6	-4.6	-4.4	3	ASL
Chloe	6	8.9	-2.9	-3	3	TA
Daisy	7	9.4	-2.4	-3.3	37	TA
Deborah	6	9.6	-3.6	-6.2	40	TA
Esme	8	9.9	-1.9	-4.9	27	TA
Esme	4	7.5	-3.5	-3.1	22	TA
Helen	6	8.9	-2.9	-3	40	TA
Holly	8	9.8	-1.8	-3.6	27	TA
Lexi	8	9.9	-1.9	-5.2	27	TA
Lucy	8	9.8	-1.8	-4	27	TA
Maddison	2	6.2	-4.2	-3.8	12	FLIN
Maddison	8	9.7	-1.7	-3.1	27	TA
Maddison	2	5.3	-3.3	-3.1	41	FLIN
Margarete	6	9.1	-3.1	-3.5	38	TA
Mary	10	6.3	3.7	3.4	9	TA
North	7	9.7	-2.7	-5.2	27	TA
North	6	9.1	-3.1	-3.6	40	TA
Paige	6	9	-3	-3.2	40	TA
Paige	3	6.5	-3.5	-3.1	22	TA
Penelope	5	9.6	-4.6	-7.2	27	TA
Penelope	2	6.2	-4.2	-3.8	22	TA
Penelope	4	7.6	-3.6	-3.2	35	TA
Rosie	7	9.5	-2.5	-3.4	40	TA
Rosie	4	7.6	-3.6	-3.2	13	ASL
Rosie	6	8.9	-2.9	-3.1	37	TA

Sally	3	7.5	-4.5	-4	22	TA
Tyler	5	9.2	-4.2	-4.8	40	TA
Tyler	9	5.7	3.3	3.1	15	ASL
Violet	5	8.7	-3.7	-3.6	37	RSL
Violet	7	9.4	-2.4	-3.3	40	TA
Willow	5	8.5	-3.5	-3.3	14	TA
Zachary	5	8.6	-3.6	-3.5	3	ASL
Zara	5	8.4	-3.4	-3.1	37	TA

Note. TA = Task achievement. FLIN = Fluency and interaction. RSL = Range of spoken language. ASL = Accuracy of spoken language.

4.3.6 Criterion-specific RSM Models

Considering the results of this first MFRM analysis, but also in order to create measures that might be more sensitive and able to detect differences in rater cognition with regards to the four criteria in the rating scale, it was decided to run separate MFRM analyses for each criterion.

Table 4.8 below summarizes the measurement statistics for all four analyses. In terms of candidate ability, candidates on average were strongest but also most widely spread out in the model for FLIN ($M_{FLIN} = 1.59$, $SD = 1.68$) and weakest in the model for RSL ($M_{RSL} = 0.55$, $SD = 1.32$). The model for the criterion TA has the lowest separation ratio (6.06) and also the least distinguished statistically separate classes of candidate ability (8.42). In terms of rater severity, raters were most widely distributed in the model for FLIN ($SD = 0.71$, separation = 3.38, strata = 4.84) and most compact for the criterion RSL ($SD = 0.52$, separation = 2.50, strata = 3.67). Finally, the criterion-specific analysis of the ratings for FLIN produced the highest number of unexpected responses ($n = 8$), compared to TA ($n = 5$), and RSL and ASL (both $n = 4$).

Table 4.8 Summary statistics of criterion-specific RSM analyses

	TA		FLIN		RSL		ASL	
Statistic	Cand.	Rater	Cand.	Rater	Cand.	Rater	Cand.	Rater
<i>M</i> (measure)	1.50	0.00	1.59	0.00	0.55	0.00	1.02	0.00
<i>SD</i> (measure)	0.97	0.56	1.68	0.71	1.32	0.52	1.15	0.64
<i>M</i> (SE)	0.16	0.18	0.21	0.20	0.19	0.19	0.17	0.19
RMSE	0.16	0.18	0.26	0.20	0.21	0.19	0.18	0.19
Adj. (true) <i>SD</i>	0.96	0.53	1.66	0.68	1.31	0.48	1.14	0.61
Chi-square	1038.5*	370.0*	1308.6*	461.2*	927.2*	276.6*	902.4*	434.6*
<i>df</i>	29	38	29	38	29	38	29	38
Separation ratio	6.06	2.98	6.48	3.38	6.24	2.50	6.48	3.25
Separation strata	8.42	4.31	8.98	4.84	8.65	3.67	8.98	4.66
Reliability of separation	0.97	0.90	0.98	0.92	0.97	0.86	0.98	0.91

Note. RMSE = root mean-square measurement error. The rater facet was centered and constrained to have a mean element measure of zero. TA = Task achievement. FLIN = Fluency and interaction. RSL = Range of spoken language. ASL = Accuracy of spoken language. * $p < .05$

Before using the resulting five measures (i.e., severity for the full model as well as severity on each criterion) as dependent variables in the subsequent analyses (see Chapter 5), approximately normal distributions were confirmed for all variables. The relationship between the various measures of severity was explored through calculating a Pearson product-moment correlation matrix for all five variables. As can be seen in Table 4.9, the severity on the criterion TA shows the most moderate relationship with the severity measures of the other criteria. ASL and RSL are both strongly related to one another and display the strongest association with the overall level of severity estimated by the full model.

Table 4.9 Pearson correlations between criterion-specific severity measures

MFRM measures	1	2	3	4
1. Severity (Full model)	-			
2. Severity (TA)	.81**	-		
3. Severity (FLIN)	.90**	.61**	-	
4. Severity (RSL)	.93**	.66**	.80**	-
5. Severity (ASL)	.93**	.63**	.82**	.89**

Note. ** Correlation is significant at the 0.01 level (2-tailed).

An analysis of the overfitting and misfitting raters for each criterion-specific MFRM model is summarized Table 4.10 and Table 4.11. Again, TA emerges as problematic with the overall highest number of misfitting raters ($n = 10$). Infit mean-squares are more sensitive to higher numbers of unexpected ratings and therefore also considered the more important index of model fit (Myford & Wolfe, 2003). In terms of infit mean-squares, only five (FLIN) or six raters (RSL, ASL) showed considerable overfit or underfit. Significant overfit (MS_w or $MS_u < 0.7$, $t > 2.0$ or $t < -2.0$) is generally considered a less serious issue as it simply indicates that the ratings observed in the sample are closer to the expected ratings than the model expects. Significant misfit (MS_w or $MS_u > 1.3$, $t > 2.0$ or $t < -2.0$), however, may indicate idiosyncratic rater behaviour that is not in line with the decision making of the other raters.

Table 4.11 lists those raters that display significant misfit or overfit in the criterion-specific analyses. As far as overfit is concerned, Daisy is a relatively extreme case in this cohort with highly overfitting mean-square indices in three criteria ($MS_{w\ TA} = 0.54$, $MS_{w\ RSL} = 0.43$, and $MS_{w\ ASL} = 0.27$). Layla and North provided highly overfitting ratings for two criteria each ($MS_{w\ TA} = 0.66$, $MS_{w\ ASL} = 0.43$, and $MS_{w\ TA} = 0.68$, $MS_{w\ RSL} = 0.46$, respectively). Rosie appears most frequently in the list of misfitting raters ($MS_{w\ TA} = 1.76$, $MS_{w\ FLIN} = 1.89$, and $MS_{w\ ASL} = 1.80$). The raters with the highest

indices of misfit are Maddison ($MS_{w\text{ TA}} = 1.88$, $MS_{w\text{ FLIN}} = 2.95$), Penelope ($MS_{w\text{ RSL}} = 2.35$, $MS_{w\text{ TA}} = 1.53$) and Chloe ($MS_{w\text{ ASL}} = 2.12$, $MS_{w\text{ TA}} = 1.41$).

Table 4.10 Number and percentage of misfitting and overfitting raters across criteria

Thresholds	TA		FLIN		RSL		ASL	
	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit
$\text{fit} < 0.7, t > \pm 2.0$	5 (13%)	5 (13%)	2 (5%)	0 (0%)	3 (8%)	3 (8%)	2 (5%)	2 (5%)
$0.7 < \text{fit} < 1.3$	29 (74%)	29 (74%)	34 (87%)	37 (95%)	33 (85%)	35 (90%)	33 (85%)	34 (87%)
$\text{fit} > 1.3, t > \pm 2.0$	5 (13%)	5 (13%)	3 (8%)	2 (5%)	3 (8%)	1 (3%)	4 (10%)	3 (8%)

Note. TA = Task achievement. FLIN = Fluency and interaction. RSL = Range of spoken language. ASL = Accuracy of spoken language.

Table 4.11 Misfitting and overfitting raters across criteria (based on infit MS)

Criterion	Overfit		Misfit	
	Rater	$MS_w (t_w)$	Rater	$MS_w (t_w)$
TA	Daisy	0.54 (-3.62)	Maddison	1.88 (4.73)
	Lexi	0.60 (-3.13)	Rosie	1.76 (4.26)
	Lucy	0.63 (-2.80)	Penelope	1.53 (3.06)
	Layla	0.66 (-2.35)	Holly	1.52 (3.05)
	North	0.68 (-2.35)	Chloe	1.41 (2.52)
FLIN	Violet	0.45 (-2.59)	Maddison	2.95 (4.92)
	Phoebe	0.54 (-2.01)	Rosie	1.89 (2.75)
			Kimberly	1.83 (2.65)
RSL	Daisy	0.43 (-2.78)	Penelope	2.35 (3.96)
	Maisie	0.44 (-2.70)	Holly	1.69 (2.35)
	North	0.46 (-2.60)	Mary	1.60 (2.07)
ASL	Daisy	0.27 (-4.05)	Chloe	2.12 (3.44)
	Layla	0.43 (-2.82)	Rosie	1.80 (2.64)
			Zachary	1.69 (2.34)
			Tyler	1.60 (2.06)

Note. MS_w = mean-square infit statistic. t_w = standardized infit statistic. Overfit was defined as having an infit mean-square of 0.70 or less and a t -score of -2 or less. Misfit was defined as having an infit mean-square of 1.30 or more and a t -score of 2 or more.

4.3.7 Summary Rater Severity

The analyses of severity, consensus and consistency as well as the MFRM analyses based on the rating scale model provide evidence which address the first subsidiary research questions: *How consistent are novice raters when using the rating scale? And does consistency vary according to scale criteria?*

CTT. The CTT-based data revealed that while many rater pairs obtained levels of consistency that may be approaching acceptable levels in operational contexts, the level of consensus between many pairs was problematic and below standard conventions.

Overall RSM analysis. The overall MFRM analysis of rater severity showed that the data fit the model reasonably well in many, if not all respects. Looking first at the examinee facet, there is evidence that the sample included a range of ability levels and that the raters were able to identify the variability through applying the rating scale, providing evidence to support the claim that raters were able to identify different levels of ability in the sample. Furthermore, the analysis revealed that raters were quite heterogeneous as far as their respective levels of severity were concerned. Several elements of the rater facet as well as the criterion facet, however, emerged as problematic for model fit. A fifth of all raters in the overall severity model displayed misfitting patterns. This may be due to several reasons. One particularly strong factor might be the criterion TA, which potentially explains a great part of the statistical anomalies of the rater measurement results. This criterion is significantly rated more leniently than the two language-related criteria (RSL, ASL) and associated with most unexpected observations or residuals in the overall model. Clearly, novice raters

appear to find it difficult to apply the descriptors of the TA dimension with the same consistency as the other three dimensions.

Criterion-specific RSM analysis. The criterion-specific series of MFRM analyses produced acceptable model fit and separation indices for each individual analysis. Model fit was weakest for the TA-specific model as almost a fourth of all raters were either overfitting or misfitting in this model. Nonetheless, all four criterion-specific models showed satisfactory model fit in terms of separation indices and residuals. The mean-square statistics based on the criterion-specific analyses also revealed raters that were particularly predictable or unpredictable in their rating patterns for criteria and provided more fine-grained information than the analysis on the overall severity of a rater.

In light of the subsidiary research question, it can be concluded that raters displayed a great variety in their severity levels, but generally were able to identify different ability levels through the descriptors of the scale. The heterogeneity of severity levels also led to considerable differences in consistency. Many rater pairings obtained acceptable levels of inter-rater consistency. The MFRM analysis of the scale dimensions, however, also revealed that individual rater inconsistency is most frequently connected to the criterion TA which appears to function differently than the other three criteria.

4.4 Rater Accuracy (CTT approach)

Rater accuracy was defined as the extent to which the ratings provided by the novice raters agreed with the reference scores modelled on the basis of the expert scores (also

see Section 2.4). This section presents how the scores provided by the participants compare to the reference scores on the basis of CTT.

As was the case in the analysis of interrater reliability in Section 4.2, four measures of consensus and consistency were used to investigate the relationships between the two sets of scores: 1) exact agreement, 2) weighted Cohen's kappa (κ), and 3) Kendall's Tau-b (τ_b). Table 4.12 presents the consensus and consistency between the rounded scores provided by novice raters and the rounded reference scores for each individual rater. The results are sorted according to their percent agreement and ranged from 60.00% for Stormi to 13.33% for Betty.

Table 4.12 Consensus and consistency of participant ratings with expert benchmarks

Rater	% exact agreement	Cohen's weighted κ	Kendall's τ_b
Stormi	60.00	.55	.59
Kimberly	56.67	.55	.60
Lexi	56.67	.62	.72
Scarlett	56.67	.61	.71
Phoebe	53.33	.56	.72
Esme	50.00	.56	.71
Jennifer	50.00	.52	.64
Alice	46.67	.54	.69
Layla	46.67	.42	.52
Violet	43.33	.41	.56
Zachary	43.33	.37	.46
Holly	40.00	.49	.71
Chloe	36.67	.41	.59
Daisy	36.67	.39	.63
Deborah	36.67	.39	.55
Eliza	36.67	.42	.62
Lucy	36.67	.49	.73
Margarete	36.67	.44	.63
Susan	36.67	.40	.55
Willow	36.67	.27	.46
Amy	33.33	.44	.71

Megan	33.33	.45	.67
Paige	33.33	.26	.34°
Zara	33.33	.34	.62
Donny	30.00	.42	.64
Dorothy	30.00	.28	.54
Hendrix	30.00	.34	.56
Nancy	30.00	.30	.53
Sally	30.00	.21°	.33°
Maisie	26.67	.29	.51
North	26.67	.36	.76
Rosie	26.67	.36	.57
Helen	23.33	.21	.51
Maddison	20.00	.29	.60
Tyler	20.00	.22°	.48
Amber	16.67	.22	.52
Mary	16.67	.19°	.52
Penelope	16.67	.20	.39°
Betty	13.33	.14°	.47
<i>M</i>	35.56	.38	.58
<i>SD</i>	12.10	.13	.11

Note. ° significant at the .05-level or not significant. All other indices are significant at the .01-level.

As can be seen from this data, the measures of consensus are persistently lower than the measures of consistency. The majority of raters attained a slight ($n = 4$; 0-20%) or fair ($n = 24$; 21-39%) level of exact agreement with the fair scores. Stormi (60.00%) is the only rater to approach a lower threshold of substantial agreement in this group. The bulk of raters either attained a fair ($n = 15$; .21-.40) or moderate ($n = 19$; .41-.60) consensus in terms of kappa. However, only two raters (Lexi and Scarlett) managed to reach substantial agreement (κ between .61-.80).

As was already the case in the analysis of interrater reliability in Section 4.2 of this chapter, Kendall's Tau-b as a measure of consistency tended to be higher than the

measures of consensus. 20 raters reached a moderate (.41-.60) or strong (.61-.80) correlation with the benchmarks.

From this analysis, raters listed further towards the top of the table (e.g., Stormi, Kimberly or Lexi) can generally be said to emerge as more accurate than the raters listed towards the bottom of this table (e.g., Betty, Penelope, Mary or Amber). One particularly interesting rater in this analysis is North, who shows relatively little consensus in terms of exact agreement (26.67%) or kappa ($\kappa = .36$) but is among the strongest raters with regards to consistency ($\tau_b = .76$), indicating that while North might be more lenient or severe than the reference scores, she ranked the performances in a relatively similar order to the reference scores. As can be seen in the cross-tabulation in Table 4.13, North agrees exactly in only 8 out of 30 performances. However, she appears systematically more severe than the reference scores and misses the reference scores by two bands in five out of 30 cases.

Table 4.13 Cross-tabulation of North's overall scores and the reference scores

Reference score	North						Row total
	5	6	7	8	9	10	
5	0	0	0	0	0	0	0
6	0	1	0	0	0	0	1
7	2	4	1	0	0	0	7
8	0	2	7	5	0	0	14
9	0	0	0	2	1	1	4
10	0	0	0	1	3	0	4
Column total	2	7	8	8	4	1	30

Note. Consensus indices are 26.67% exact agreement and .36 Cohen's weighted kappa. Consistency index is $\tau_b = .76$.

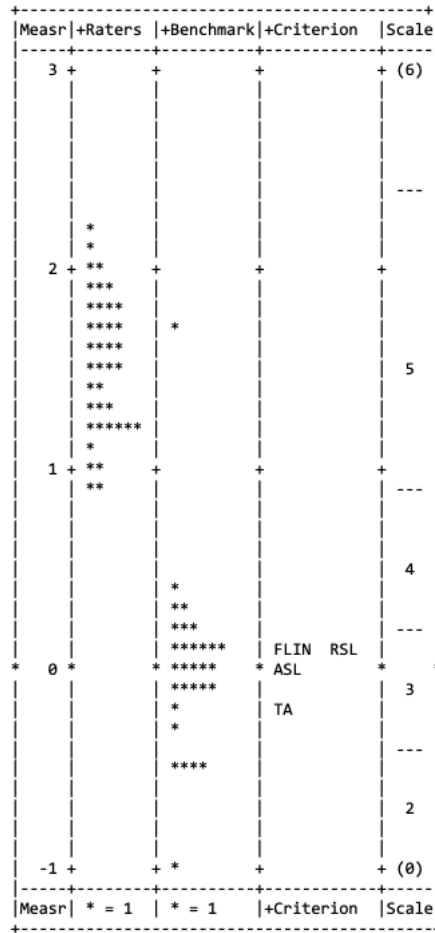
4.5 Rater Accuracy (MFRM)

The MFRM analysis of the accuracy scores was based on the polytomous rater accuracy model (RAM, Engelhard 1996) and included three facets: raters, benchmarks (i.e., reference scores), criteria. The models are based on accuracy scores, which are calculated by the differences between the rating provided by the raters and the reference score. The closer a rater's decision was to the reference score the higher the score they were awarded. This model is, thus, comparable to providing partial credit scores.

4.5.1 Wright Map

As the focus of this analysis was the accuracy of the individual raters, the rater facet was floated and the benchmark and criterion facets were centred. As indicated in the column headings of the Wright Map (Figure 4.2), all three criteria are positively oriented; therefore, all elements (raters, benchmarks and criteria) displayed further up the map, can be considered to be rated more accurately and conversely, all elements displayed further down were rated less accurately. Several particular features about this data set can already be glimpsed through the Wright Map: 1) the raters all appear rather closely bunched together and most of them are within the range of one logit; 2) there are at least two performances (see benchmarks column) that are considerably easier or more difficult to rate accurately than the majority of the performances; and 3) rating the criterion TA accurately appears to be more difficult than rating the other three criteria.

Figure 4.2 Measurement rulers for RAM analysis



4.5.2 Group-level Statistics

The summary statistics for the three-facet RAM are presented in Table 4.14. The descriptive statistics with quite low values for standard deviation for all three facets indicate a lower degree of dispersion than, in comparison, for the RSM analyses presented in Section 4.3.2. The precision of the estimates (MSE) ranges from 0.03 for the criterion facet to 0.11 for the rater facet and appear to be acceptable. The separation ratio is quite low for all three facets and indicates lower levels of dispersion; the spread of estimates in the samples is narrow in relation to the precision of measurement. Finally, all three facets contain several statistically significant levels of accuracy; in the case of raters there are four strata, for the benchmarked performances six and for

the criterion five. The separation statistics are lower in the accuracy score model (RAM) than in the severity model (RSM), suggesting that there is less dispersion in the data set.

Table 4.14 Summary statistic for rater accuracy analysis

Statistic	Rater	Benchmark	Criterion
<i>M</i> (measure)	1.52	0.00 ^a	0.00 ^a
<i>SD</i> (measure)	0.33	0.43	0.13
<i>M</i> (<i>SE</i>)	0.11	0.09	0.03
RMSE	0.11	0.10	0.03
Adj. (true) <i>SD</i>	0.31	0.42	0.12
Chi-square	372.0*	496.4*	61.1*
<i>df</i>	38	29	3
Separation ratio	2.87	4.34	3.65
Separation strata	4.15	6.12	5.20
Reliability of separation	.89	.95	.93

Note. RMSE = root mean-square measurement error. ^aThe benchmark and criterion facets were centred and constrained to have a mean element measure of zero. * $p < .05$

4.5.3 Rater Measurement

Table 4.15 below ranks all raters according to their accuracy estimates from most to least accurate. The mean accuracy measure for raters was 1.52 logit ($SD = 0.33$). With a logit of 2.15 ($SD = 0.13$), Stormi is the most accurate rater while Amber is the least accurate rater (0.93 logit, $SD = 0.09$). Stormi obtained an observed accuracy score of 643 and Amber a score of 533. Applying the same infit and outfit mean-square thresholds as in the rater severity analysis (Section 4.3.3), only two raters display significant misfit (Penelope $MS_w = 1.32$, $t_w = 2.08$, $MS_u = 1.46$, $t_u = 2.88$; Maddison $MS_w = 1.34$, $t_w = 2.18$, $MS_u = 1.39$, $t_u = 2.46$). A correlation analysis comparing the rating accuracy measure with the traditional measures of rating accuracy shows that

the Rasch estimates correlate stronger with the percentage of exact agreement ($r = .807, p < .001$) and kappa ($r = .809, p < .001$) than with tau-b ($r = .518, p = .001$).

Table 4.15 RAM estimates and fit statistics for rater facet

Rater	Total score	Obs. <i>M</i>	Measure (Accuracy)	<i>SE</i>	<i>MS_w</i>	<i>t_w</i>	<i>MS_u</i>	<i>t_u</i>
Stormi	643	5.36	2.15	0.13	0.90	-0.64	0.89	-0.68
Lexi	638	5.32	2.07	0.13	0.83	-1.14	0.93	-0.45
Alice	632	5.27	1.98	0.12	0.93	-0.41	0.95	-0.26
Jennifer	632	5.27	1.98	0.12	0.74	-1.87	0.75	-1.82
Scarlett	629	5.24	1.94	0.12	0.94	-0.36	0.86	-0.92
Lucy	625	5.21	1.88	0.12	0.89	-0.67	0.95	-0.29
Amy	623	5.19	1.85	0.12	0.98	-0.10	0.97	-0.14
Layla	622	5.18	1.84	0.12	0.95	-0.28	0.89	-0.73
Megan	622	5.18	1.84	0.12	0.90	-0.64	0.90	-0.63
Violet	617	5.14	1.77	0.11	1.10	0.68	1.07	0.53
Daisy	616	5.13	1.76	0.11	1.01	0.14	1.02	0.21
Deborah	612	5.10	1.70	0.11	1.02	0.19	1.00	0.03
Eliza	612	5.10	1.70	0.11	0.89	-0.68	0.87	-0.90
Margarete	611	5.09	1.69	0.11	0.83	-1.14	0.82	-1.23
Susan	609	5.07	1.67	0.11	0.97	-0.15	0.87	-0.90
Phoebe	607	5.06	1.64	0.11	0.95	-0.29	0.93	-0.42
Chloe	604	5.03	1.61	0.11	1.21	1.37	1.26	1.67
Donny	602	5.02	1.58	0.11	0.93	-0.45	0.92	-0.48
Zachary	601	5.01	1.57	0.11	1.09	0.64	1.05	0.37
Willow	598	4.98	1.54	0.11	1.17	1.12	1.09	0.66
Kimberly	597	4.97	1.53	0.11	0.87	-0.86	0.86	-0.92
Sally	597	4.97	1.53	0.11	1.14	0.91	1.08	0.56
Esme	594	4.95	1.49	0.11	0.89	-0.73	0.92	-0.52
North	589	4.91	1.44	0.10	0.97	-0.13	1.09	0.66
Hendrix	587	4.89	1.42	0.10	0.83	-1.16	0.83	-1.15
Holly	579	4.82	1.33	0.10	1.06	0.44	1.12	0.80
Rosie	577	4.81	1.31	0.10	1.07	0.51	1.06	0.48
Zara	576	4.80	1.30	0.10	1.21	1.34	1.19	1.29
Maisie	569	4.74	1.24	0.10	0.77	-1.66	0.80	-1.43
Paige	568	4.73	1.23	0.10	1.23	1.48	1.20	1.35
Dorothy	566	4.72	1.21	0.10	0.89	-0.71	0.91	-0.55
Helen	566	4.72	1.21	0.10	1.05	0.38	1.03	0.23

Tyler	564	4.70	1.19	0.10	0.96	-0.26	0.99	-0.04
Nancy	561	4.68	1.16	0.09	0.93	-0.42	0.88	-0.81
Maddison	549	4.57	1.06	0.09	1.34	2.18	1.39	2.46
Mary	546	4.55	1.03	0.09	0.92	-0.52	0.94	-0.41
Betty	544	4.53	1.02	0.09	0.84	-1.14	0.85	-1.05
Penelope	535	4.46	0.94	0.09	1.32	2.08	1.46	2.88
Amber	533	4.44	0.93	0.09	1.01	0.15	0.91	-0.57
<i>M</i>	593.60	4.95	1.52	0.11	0.99	-0.10	0.99	-0.10
<i>SD</i>	29.60	0.25	0.33	0.01	0.14	0.90	0.15	1.00

Note. MS_w = mean-square infit statistic. t_w = standardized infit statistic. MS_u = mean-square outfit statistic. t_u = standardized outfit statistic.

4.5.4 Criterion Measurement

Concerning the measures of the criterion facet, Table 4.16 shows that there are considerable differences between the level of accuracy raters could achieve for the criterion TA (logit = -0.21, SE = 0.03) in comparison to the other three criteria, FLIN (logit = 0.12, SE = 0.04), RSL (logit = 0.08, SE = .03) and ASL (logit = 0.01, SE = 0.03). Raters were the least accurate when rating the criterion TA and the most accurate with the criterion FLIN. This is also visible in the variable map, where the criterion TA is discernibly separate from the other three criteria. Not surprisingly, the Wald statistic confirms statistically significant differences in accuracy between TA and each of the other three criteria is; $t_{TA,ASL}(1169) = 5.19, p = .001$, $t_{TA,RSL}(1169) = 6.84, p = .001$, $t_{TA,FLIN}(1169) = 7.78, p = .001$. Furthermore, the difference in terms of accuracy between FLIN and ASL of 0.11 logits is also significant; $t_{FLIN,ASL}(1169) = 2.59, p = .01$. The accuracy difference between RSL and FLIN is too small to be significant; $t_{FLIN,RSL}(1169) = 0.94, ns$.

Table 4.16 Accuracy measures and fit statistics for the criterion facet

Criterion	Total score	Obs. <i>M</i>	Measure (Accuracy)	<i>SE</i>	<i>MS_w</i>	<i>t_w</i>	<i>MS_u</i>	<i>t_u</i>
FLIN	5898	5.04	0.12	0.04	0.95	-0.92	0.94	-1.19
RSL	5859	5.01	0.08	0.03	0.92	-1.63	0.89	-2.37
ASL	5799	4.96	0.01	0.03	0.99	-0.24	0.97	-0.61
TA	5596	4.78	-0.21	0.03	1.11	2.15	1.15	2.99

Note. *MS_w* = mean-square infit statistic. *t_w* = standardized infit statistic. *MS_u* = mean-square outfit statistic. *t_u* = standardized outfit statistic. TA = Task achievement. FLIN = Fluency and interaction. RSL = Range of spoken language. ASL = Accuracy of spoken language. Estimates based on 1,170 observations.

4.5.5 Performance Measurement

The measurement report for the benchmarks facet reveals the relative difficulty of rating specific performances accurately (Table 4.17). Performance 27 was by far the easiest to rate accurately (logit = 1.69, *SE* = 0.17), while performance 15 was the most difficult to rate accurately (logit = -0.96, *SE* = 0.07). Six out of the 30 performances included in the accuracy analysis display significant misfit or overfit. Performances 38 and 21 showed comparatively lower means-square statistics and negative *t*-values which are indicators of a tendency for overfit (*MS_w* = 0.70, *t_w* = -2.47; *MS_w* = 0.69, *t_w* = -2.62, respectively). The other four performances (27, 40, 35 and 37) produced significant mean-square measures between 1.33 and 1.91 all of which indicate misfit and are less useful for fitting the model to the data. A comparison of the misfitting and underfitting performances with the ability estimates based on the RSM model shows that most of these performances are towards the upper end of the ability scale; performances 27, 40 and 38 are the top three performers included in the sample. Performance 37 ranked 5th and 35 ranked 7th. Performance 21, on the other hand, was the second weakest speaker included in the sample.

Table 4.17 Accuracy estimates and fit statistics for underfitting speakers

Speakers	Total score	Obs. <i>M</i>	Measure (Accuracy)	<i>SE</i>	<i>MS_w</i>	<i>t_w</i>	<i>MS_u</i>	<i>t_u</i>
27	895	5.74	1.69	0.17	1.67	3.47	1.37	2.01
38	817	5.24	0.39	0.11	0.7	-2.47	0.69	-2.70
40	809	5.19	0.30	0.10	1.91	5.54	1.74	4.78
21	748	4.79	-0.24	0.09	0.69	-2.62	0.69	-2.66
35	713	4.57	-0.49	0.08	1.33	2.38	1.23	1.75
37	709	4.54	-0.51	0.08	1.41	2.91	1.35	2.54

Note. *MS_w* = mean-square infit statistic. *t_w* = standardized infit statistic. *MS_u* = mean-square outfit statistic. *t_u* = standardized outfit statistic.

4.5.6 Global Model Fit

Overall, there were 31 observations with unexpected responses. About half of the residuals ($n = 16$) were on the criterion TA which fits with the criterion measurement data discussed in 4.3.4 that already indicated that TA was the most difficult criterion for the raters to rate accurately. The other residuals were spread quite evenly among the other three criterion elements FLIN ($n = 6$), ASL ($n = 6$) and RSL ($n = 3$). The 31 residuals were spread over 24 different raters, whereby five raters produced two unexpected observations (Chloe, North, Tyler, Paige, Sally) and only one rater produced three (Holly). Overall, with 0.66% of residuals with a difference of ≥ 3 between expected and observed score, global model fit is satisfactory.

4.5.7 Rater-criterion Interactions

An exploratory bias analysis was conducted to detect systematic interactions affecting rating accuracy. The following three interactions were modelled: raters with certain benchmarks (1,170 interaction terms), raters with certain criteria (156 interaction terms), and raters with certain benchmark performances with certain criteria (4,680 interaction terms). While there is a considerable percentage of potentially problematic

pairings with large t -values for both, the rater-benchmark (7.18%) and rater-criterion interaction analysis (9.62%), only the rater-criterion-interaction warrants closer scrutiny based on the proportion of statistically significant interactions (8.97%).

Table 4.18 Summary statistics for the exploratory interaction analysis

Statistic	Rater x Benchmark	Rater x Criterion	Rater x Benchmark x Criterion
N	1,170	156	4,680
% large t -values	7.18	9.62	2.28
% sign. t -values	1.2	8.97	0
Min- t (df)	-14.91	-128.47	-3.69 (1)
Max- t (df)	2.32 (3)	2.82** (29)	1.57 (1)
M	0	-0.02	-0.14
SD	1.18	1.18	0.78

Note. N = number of bias terms. Percentage of absolute t -values equal or greater than 2. Percentage of t -values statistically significant at * $p < .05$. ** $p < .01$

Figure 4.3 provides an overview of all rater-criterion interactions identified through the bias analysis. Fourteen out of the 156 possible bias terms show a significant interaction between elements of the rater and criterion facet with a t -value below -2 or above +2. As can be seen from comparing this figure with the rater accuracy estimates (see also Table 4.19), bias effects emerged with capable and accurate raters with high overall accuracy measures (e.g., Violet, 1.77 logit) as well as raters with low overall accuracy measures (e.g., Penelope, 0.94 logit). In nine of the fourteen cases, the bias measure had a positive value indicating that the raters were more accurate in rating that particular criterion than expected by the model. Slightly less than half of these cases ($n = 4$) concern the criterion TA, which was the most difficult criterion to rate accurately and displayed more bias interactions ($n = 5$) than the three other criteria (3 interactions each).

In five cases (Daisy, Helen, Mary, Sally and Violet) the bias patterns reveal significant unexpected ratings for two criteria. Daisy, an overall quite accurate rater (1.76 logit),

displayed highly accurate ratings for both ASL and RSL, but produced less accurate ratings for the criterion TA and significantly less accurate ratings for FLIN. Sally, who is overall an average rater in terms of rating accuracy (1.53 logit), achieved a decisively lower accuracy score for the criterion TA than expected (observed score 122 versus expected score 144.44).

Figure 4.3 Bias diagram of interactions between raters and criteria (statistically significant interactions are circled)

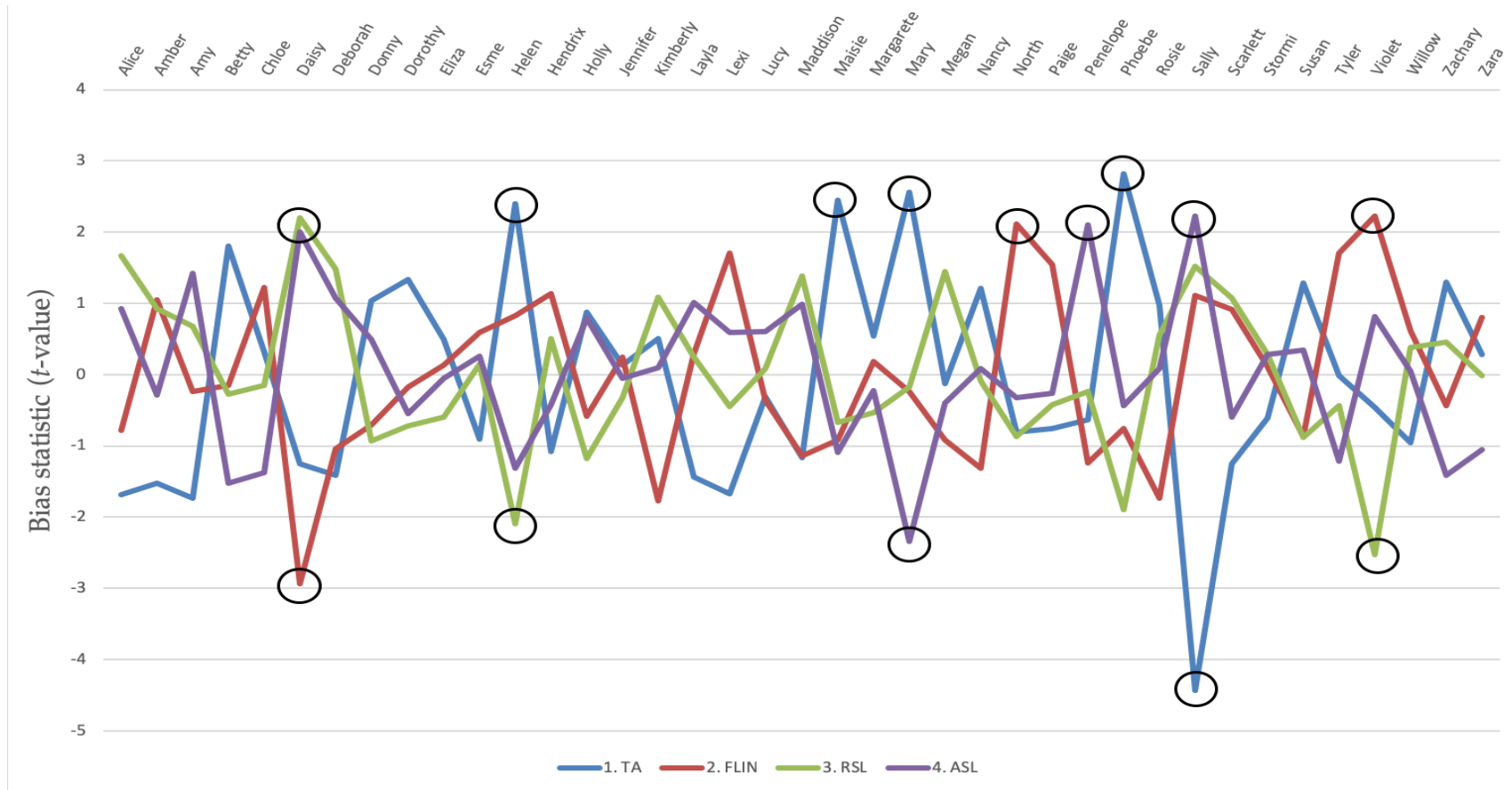


Table 4.19 Significant rater-criterion interactions

Rater	Accuracy measure	Criterion	Accuracy measure	Obs.	Exp.	Bias Measure	SE	<i>t</i>	<i>p</i>
Phoebe	1.64	TA	-.21	161	147.28	.73	.26	2.82	.009
Violet	1.77	FLIN	.12	166	156.61	.66	.30	2.22	.034
Daisy	1.76	RSL	.08	165	155.53	.63	.29	2.20	.036
Sally	1.53	ASL	.01	160	149.52	.57	.25	2.23	.034
North	1.44	FLIN	.12	160	150.15	.54	.25	2.11	.043
Maisie	1.24	TA	-.21	150	136.48	.53	.22	2.45	.021
Helen	1.21	TA	-.21	149	135.62	.51	.21	2.40	.023
Mary	1.03	TA	-.21	145	129.93	.51	.20	2.56	.016
Penelope	0.94	ASL	.01	146	134.14	.43	.20	2.10	.045
Helen	1.21	RSL	.08	133	143.66	-.37	.18	-2.10	.044
Mary	1.03	ASL	.01	124	136.87	-.38	.16	-2.34	.026
Violet	1.77	RSL	.08	145	155.77	-.51	.20	-2.53	.017
Daisy	1.76	FLIN	.12	144	156.37	-.58	.20	-2.93	.007
Sally	1.53	TA	-.21	122	144.44	-.71	.16	-4.43	.000

Note. TA = Task achievement. FLIN = Fluency and interaction. RSL = Range of spoken language. ASL = Accuracy of spoken language.

4.5.8 Criterion-specific RAM Models

As was the case in the RSM-based analysis of rater severity and fit, the different behaviour of the criterion element TA suggests that the assumption of unidimensionality of the latent variable is not fully met by the data. The estimates related to the TA element indicate that this element is significantly more difficult to rate accurately than the other three elements of the criterion facet. Furthermore, about a quarter of raters ($n = 9$) show bias towards at least one criterion and 16 unexpected observations with a residual value of 3 or higher were associated with the criterion TA. To produce more sensitive measures, a separate RAM analysis for each criterion was carried out. The summary results for these separate criterion-specific analyses are presented in Table 4.20.

Table 4.20 Summary statistics criterion-specific RAM analyses

Statistic	TA		FLIN		RSL		ASL	
	Raters	RS	Raters	RS	Raters	RS	Raters	RS
<i>M</i> (measure)	1.47	0.00 ^a	1.85	0.00 ^a	1.58	0.00 ^a	1.65	0.00 ^a
<i>SD</i> (measure)	0.38	0.56	0.50	0.91	0.50	0.75	0.50	0.53
<i>M</i> (SE)	0.21	0.19	0.24	0.15	0.24	0.22	0.23	0.20
RMSE	0.21	0.19	0.24	0.28	0.24	0.24	0.23	0.21
Adj. (true) <i>SD</i>	0.32	0.53	0.44	0.86	0.44	0.71	0.45	0.49
Chi-square	145.1*	359.3*	162.7*	174.1*	160.1*	148.8*	190.5*	130.0*
<i>df</i>	38	29	38	29	38	29	38	29
Separation ratio	1.52	2.78	1.79	3.11	1.83	2.97	1.93	2.36
Separation strata	2.37	4.04	2.72	4.49	2.77	4.30	2.90	3.48
Reliability of separation	.70	.89	.76	.91	.77	.90	.79	.85

Note. RMSE = root mean-square measurement error. The rater facet was centred and constrained to have a mean element measure of zero. RS = reference score. TA = Task achievement. FLIN = Fluency and interaction. RSL = Range of spoken language. ASL = Accuracy of spoken language. * $p < .05$

The separate MFRM analyses modelled for each criterion produced five variables. The intercorrelation matrix in Table 4.21 summarizes the relationships between these five variables. Accuracy for the criterion TA displayed more distinct properties in that it showed the weakest relationship of all criterion-based variables with the measure based on the full model ($r = .63, p < .001$.) and no significant correlation to the accuracy measures based on the other criteria. Thus, being accurate when assessing TA appears to contribute less to the overall accuracy of a rater than accuracy on any of the other three criteria. Accuracy in assessing TA is also only weakly associated with rating the other criteria. The strongest association among the criterion-based variables is between the two language related criteria ASL and RSL ($r = .83, p < .001$). Being accurate on these two criteria also contributed most to overall accuracy as these two variables correlated strongly with the accuracy measure based on the full model (RSL $r = .85, p < .001$; ASL $r = .88, p < .001$).

Table 4.21 Pearson correlations between criterion-specific accuracy measures

MFRM measures	1	2	3	4
1. Accuracy (Full model)	-			
2. Accuracy (TA)	.63**	-		
3. Accuracy (FLIN)	.80**	.37*	-	
4. Accuracy (RSL)	.85**	.30	.53**	-
5. Accuracy (ASL)	.88**	.33	.62**	.83**
<i>Note. TA = Task achievement. FLIN = Fluency and interaction. RSL = Range of spoken language. ASL = Accuracy of spoken language. ** Correlation is significant at the 0.01 level (2-tailed). * Correlation is significant at the .05-level (2-tailed).</i>				

As a result of the separate criterion-specific analyses, it is also possible to identify particularly accurate and inaccurate raters for each criterion. Table 4.22 and Table 4.23 present the most and least accurate raters for each of the four criteria. The most

and least accurate raters were determined by isolating raters within the first standard deviation from the maximum logit score or one standard deviation from the minimum logit score, respectively. Despite providing measures towards the upper or lower end of the spectrum, the majority of raters included in these two lists provided mean-square fit statistics within a narrowly defined fit range of an upper-limit threshold at 1.30 and lower limit threshold at .70. A few raters, however, produced misfitting or overfitting scores. Jennifer (TA accurate), Daisy (RSL accurate), Sally (TA inaccurate), Mary (FLIN inaccurate) and Masie (FLIN, RSL and ASL inaccurate) show tendencies of overfit. Raters Penelope (TA and RSL inaccurate), Maddison (FLIN inaccurate), Paige (RSL inaccurate) and Amber (ASL inaccurate), on the other hand, display significant tendencies of misfit.

After completing the RAM analysis, five dependent variables were formed to be used for subsequent analyses. The continuous variables were based directly on the logit scores as derived from the full model as well as the four separate, criterion-specific models. Due to the nature of the transformed accuracy scores, RAM produces more muted data. As there were overall only seven instances of overfit and five instances of misfit out of 156 measures, no meaningful groups or variables could be created based on these results and the aspect of fit in the context of accuracy and the RAM was disregarded for later analyses.

Table 4.22 Most accurate raters by criterion

Criterion	Total Score	Measure (accuracy)	SE	MS_w	t_w	MS_u	t_u	Rater
TA	161	2.39	.27	0.93	-0.12	0.99	0.07	Phoebe
	155	2.00	.24	0.64	-1.23	0.61	-1.47	Jennifer
	155	2.00	.24	0.89	-0.25	0.88	-0.35	Stormi
FLIN	168	3.03	.32	1.06	0.31	1.01	0.18	Lexi
	166	2.84	.30	0.88	-0.33	0.86	-0.16	Violet
	163	2.58	.28	1.01	0.13	1.00	0.13	Scarlett
	163	2.58	.28	0.92	-0.19	0.89	-0.14	Stormi
RSL	166	2.61	.31	0.90	-0.24	0.94	-0.02	Alice
	165	2.52	.30	0.67	-1.16	0.68	-0.85	Daisy
	163	2.35	.28	0.97	-0.01	0.94	-0.06	Megan
	163	2.35	.28	0.72	-0.97	0.69	-0.86	Scarlett
	163	2.35	.28	1.06	0.31	0.93	-0.08	Stormi
	161	2.20	.27	0.95	-0.07	0.95	-0.03	Deborah
ASL	160	2.12	.27	1.17	0.67	1.19	0.65	Amy
	163	2.47	.28	0.96	-0.06	1.05	0.25	Daisy
	162	2.39	.27	0.72	-1.01	1.00	0.12	Alice
	162	2.39	.27	0.96	-0.06	0.99	0.08	Amy
	162	2.39	.27	0.87	-0.38	0.85	-0.44	Lexi
	162	2.39	.27	1.02	0.17	0.97	-0.01	Stormi
	160	2.25	.26	0.71	-1.04	0.68	-1.15	Layla
	160	2.25	.26	1.14	0.55	1.32	1.09	Sally
	159	2.18	.26	0.95	-0.08	0.86	-0.41	Lucy
	158	2.12	.25	1.27	0.99	1.23	0.84	Deborah
	158	2.12	.25	1.08	0.37	1.08	0.38	Jennifer
	158	2.12	.25	0.89	-0.31	0.86	-0.42	Violet

Note. MS_w = mean-square infit statistic. t_w = standardized infit statistic. MS_u = mean-square outfit statistic. t_u = standardized outfit statistic. TA = Task achievement. FLIN = Fluency and interaction. RSL = Range of spoken language. ASL = Accuracy of spoken language. Raters were considered most accurate if their ratings were within one standard deviation of the maximum accuracy measure.

Table 4.23 Least accurate raters by criterion

Criterion	Total Score	Measure (accuracy)	SE	MS_w	t_w	MS_u	t_u	Rater
TA	117	0.55	.16	1.01	0.13	0.93	-0.17	Amber
	122	0.69	.17	0.6	-1.67	0.52	-2.04	Sally
	123	0.72	.17	1.89	2.74	2.01	2.92	Penelope
	124	0.75	.17	1.08	0.38	1.13	0.55	Maddison
FLIN	131	0.86	.20	0.89	-0.34	1.33	1.11	Penelope
	135	1.02	.20	1.91	2.72	1.76	2.16	Maddison
	137	1.11	.21	0.94	-0.14	0.92	-0.17	Nancy
	139	1.19	.21	0.99	0.05	1.01	0.12	Betty
	139	1.19	.21	0.64	-1.39	0.6	-1.41	Mary
	139	1.19	.21	1.16	0.64	1.13	0.5	Rosie
	141	1.28	.21	0.68	-1.19	0.66	-1.12	Maisie
RSL	133	0.76	.19	1.01	0.14	0.94	-0.12	Helen
	135	0.84	.20	1.44	1.52	1.55	1.77	Penelope
	137	0.92	.20	0.87	-0.42	0.79	-0.7	Betty
	138	0.96	.20	1.11	0.48	1.13	0.53	Mary
	140	1.04	.21	1.22	0.84	1.18	0.69	Dorothy
	141	1.08	.21	0.87	-0.4	0.72	-0.98	Amber
	141	1.08	.21	1.03	0.2	0.92	-0.18	Holly
	141	1.08	.21	0.66	-1.3	0.69	-1.1	Maisie
	141	1.08	.21	0.87	-0.39	0.85	-0.43	Tyler
	142	1.13	.21	0.93	-0.16	0.9	-0.27	Nancy
ASL	142	1.13	.21	1.7	2.17	1.59	1.8	Paige
	124	0.59	.18	1.06	0.31	1.05	0.26	Mary
	128	0.73	.19	0.77	-0.91	0.79	-0.83	Betty
	132	0.88	.19	1.40	1.48	1.33	1.23	Amber
	135	0.99	.20	0.94	-0.13	0.91	-0.25	Helen
	135	0.99	.20	0.90	-0.29	0.93	-0.17	Tyler
	137	1.07	.20	0.65	-1.42	0.65	-1.39	Maisie

Note. MS_w = mean-square infit statistic. t_w = standardized infit statistic. MS_u = mean-square outfit statistic. t_u = standardized outfit statistic. TA = Task achievement. FLIN = Fluency and interaction. RSL = Range of spoken language. ASL = Accuracy of spoken language. Raters were considered least accurate if their ratings were within one standard deviation of the minimum accuracy measure.

4.5.9 Summary Rater Accuracy

The results presented in the previous sections (4.5.1 to 4.5.6) address the second subsidiary research question investigated in this study: *How accurate are novice raters when using the scale? And does accuracy vary according to scale criteria?*

CTT. Few raters managed to obtain adequate consensus with the reference scores. In terms of consistency, half of the cohort reached moderate correlations in ranking the performances ($\tau_b = .41$ or higher).

Overall RAM analysis. The MFRM analysis based on the rater accuracy scores (RAM) produced a model with improved fit as compared to the RSM model. The analysis revealed at least four statistically separate levels of rater accuracy. Accuracy estimates ranged from 2.15 (Stormi) to 0.93 (Amber) ($M = 1.52$, $SD = 0.33$). The criterion TA was significantly more difficult to rate accurately and also produced slightly more unexpected observations. The performance facet showed that particularly performances toward the upper end of the ability scale tended to produce statistically significant misfit.

Bias analysis. An exploratory investigation of facet interactions revealed that there was a considerable proportion of statistically significant rater-criterion interactions (8.92%) which occurred across all four criteria, but most frequently ($n = 5$) with the criterion TA. Furthermore, five out of 39 raters were found to display accuracy interactions with two criteria.

Criterion-specific RAM analysis. Correlating the four criterion-specific accuracy estimates with overall accuracy again confirmed differential patterns for the

criterion TA. Being accurate in rating TA appears to be less associated with overall accuracy or accuracy in rating the other three criteria.

In sum, there is some evidence to support the assumption that this group of novice raters was able to use the Austrian rating scale in terms of *consistency* with the reference scores. *Consensus* with the reference scores, however, would not meet the conventional standards. There were great differences among the raters with four distinct levels of accuracy evident in the overall RAM model fit. There was evidence that TA was more difficult to rate accurately which was supported by the criterion-specific analyses as well as exploratory bias analysis.

4.6 Perception Data

At several points throughout the experiment, additional data concerning the participants' experience of rating the performances was collected. At four points of the experiment (i.e., after each quarter) participants were asked how difficult they found rating the set of performances they had just seen (1 = very difficult, 5 = very easy) and how confident they were in their ratings (1 = not confident, 5 = very confident). After each rating session, raters also judged the difficulty of each criterion (1 = very difficult, 4 = very easy) and provide a justification.

4.6.1 General Confidence in Rating Decisions

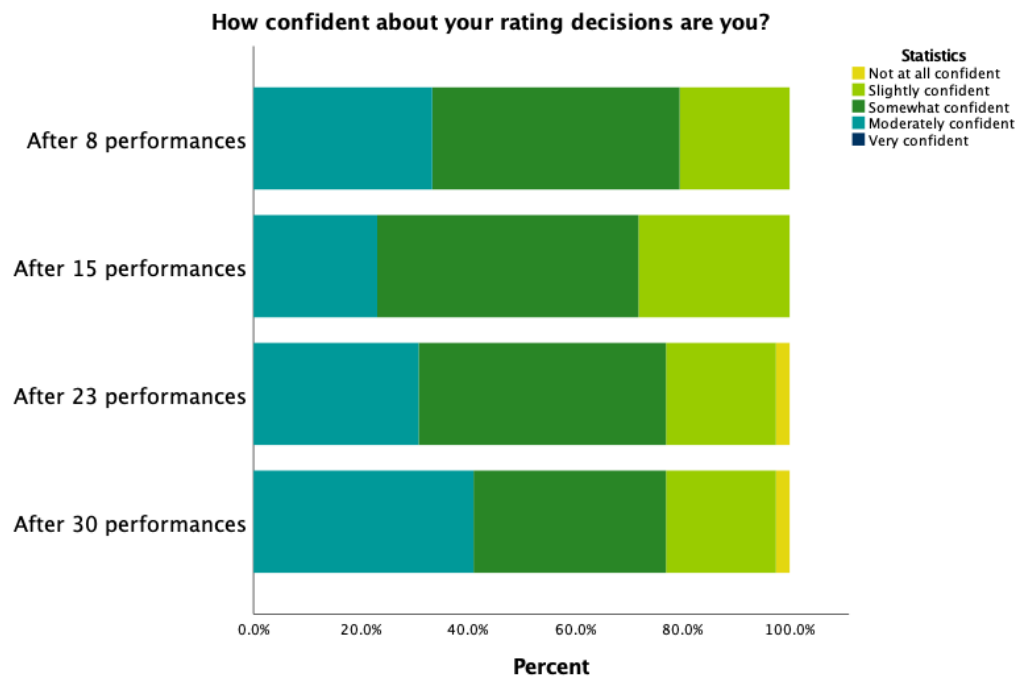
Table 4.24 provides an overview of the descriptive statistics on participants' confidence regarding their rating decisions. None of the participants considered themselves to be "very confident" about their rating decisions at the end of the experiment. However, there was also only one participant who was "not at all

confident” after the second rating session. As can be seen in Figure 4.4, the proportion of moderately confident raters is largest at the end of the second session (after 30 performances), followed by the middle point of the first rating session (after 15 performances). Raters overall seemed least confident at the end of the first session, i.e., after 8 performances. At the beginning of both rating sessions, participant responses were similar in terms of frequencies, but there were greater differences between the data at the end of each rating session. The greatest shift appears to have been between the third and the fourth Likert option, with more participants expressing “moderate” confidence toward the end of the experiment than in the beginning. Around three quarters of participants (76.9%) were either somewhat (35.90%) or moderately (41.00%) confident about their ratings at the end of the experiment.

Table 4.24 Self-reported confidence about rating decisions (percentage row-wise)

	Not at all confident		Slightly confident		Somewhat confident		Moderately confident		Very confident	
	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
Session 1										
After 8	0	0	8	20.50	18	46.20	13	33.30	0	0
After 15	0	0	11	28.20	19	48.70	9	23.10	0	0
Session 2										
After 23	1	2.60	8	20.50	18	46.20	12	30.80	0	0
After 30	1	2.60	8	20.50	14	35.90	16	41.00	0	0
Overall	2	1.28	35	22.44	69	44.23	50	32.10	0	0

Figure 4.4 Confidence in rating decisions



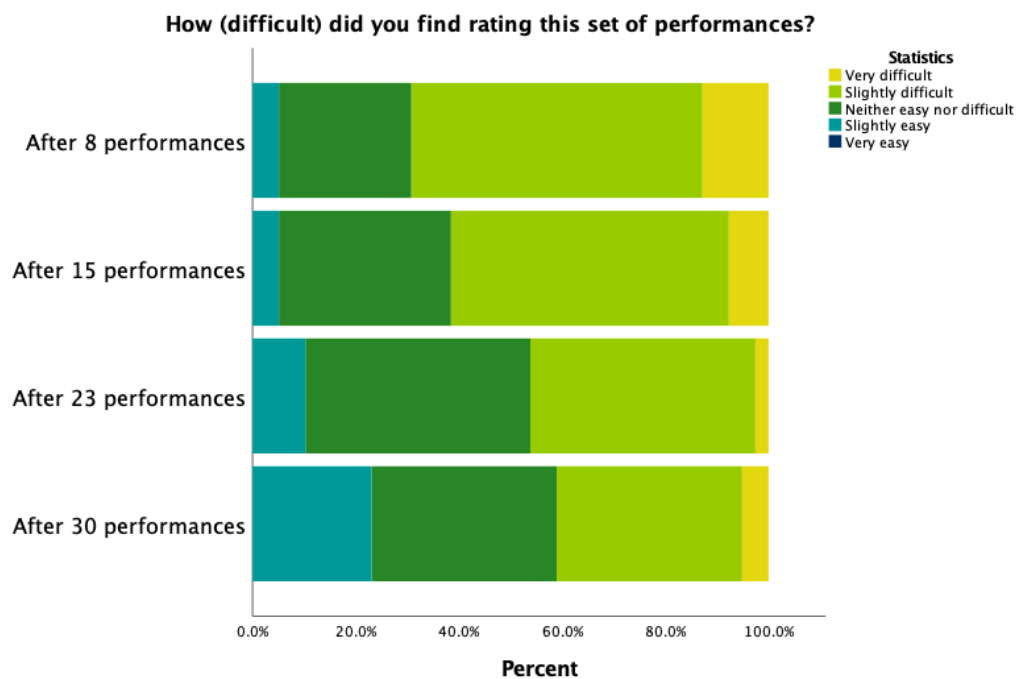
4.6.2 General Perceived Difficulty of Rating

Same as with confidence in rating decisions, raters were surveyed about how difficult they considered the rating task (see Table 4.25). Again, no participant felt that rating the performances was “very easy”. On the other hand, the number of participants who felt that rating was “very difficult” dropped from five (12.80%) after the first half of Session 1 to two (5.10%) towards the end of the experiment. There is also a clear decreasing trend in the category “slightly difficult” paired with a continuous increase of participants that felt rating was “slightly easy”. At the end of the experiment, 59% of participants perceived rating to be neither easy nor difficult or even slightly easy (see also Figure 4.5).

Table 4.25 Perceived difficulty of rating the performance (percentage row-wise)

	Very difficult		Slightly difficult		Neither easy nor difficult		Slightly easy		Very easy	
	N	%	N	%	N	%	N	%	N	%
Session 1										
After 8	5	12.80	22	56.40	10	25.60	2	5.10	0	0.00
After 15	3	7.70	21	53.80	13	33.30	2	5.10	0	0.00
Session 2										
After 23	1	2.60	17	43.60	17	43.60	4	10.30	0	0.00
After 30	2	5.10	14	35.90	14	35.90	9	23.10	0	0.00
Overall	11	7.05	74	47.44	54	34.62	17	10.90	0	0.0

Figure 4.5 Difficulty of rating



A Spearman correlation was run to investigate whether there was a relationship between confidence in rating decisions and perceived difficulty of rating and rating quality. The analysis revealed no significant correlations after *p*-value corrections between accuracy and severity and either confidence or perceived difficulty. However, there was a moderate statistically significant correlation between

perceived difficulty and confidence ($r_s = .49, p < .01$) indicating that more confident raters also felt that rating was easier and vice versa.

4.6.3 Perceived Difficulty of Using the Rating Scale

In this section the results from the Likert-type items investigating the perceived difficulty of rating each criterion will be presented. This item was placed at the end of the online rating form of each rating session and raters had to justify their choices with a short statement.

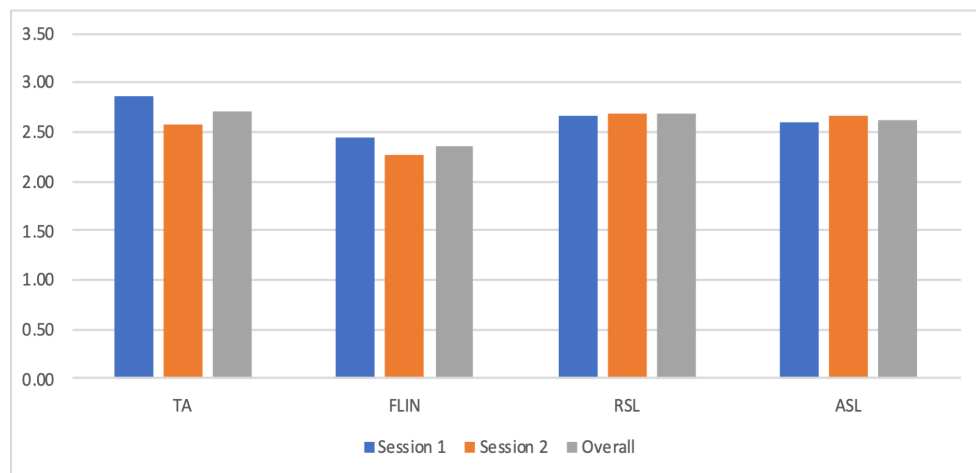
Table 4.26 summarizes the frequency counts and percentage of all four criteria over the two rating sessions. On an overall level, TA was most often judged to be *very difficult* to rate (15.6 %) while ASL and RSL were most frequently judged to be *difficult* (ASL 52.6%, RSL 62.8 %). FLIN emerges as an easier criterion (*easy* = 64.1 %). The response option *very easy* was hardly ever chosen by raters.

Figure 4.6 below illustrates and compares the mean difficulty of rating each criterion per session and overall. On average, the criterion TA was judged to be the most difficult ($M = 2.71$). However, it is also the criterion with the largest difference or, in this case, improvement between the two rating sessions. The mean value for Session 2 ($M = 2.56$) lies below the mean values for both, RSL and ASL ($M = 2.69$ and $M = 2.67$, respectively). The criterion FLIN was the easiest overall and there is a perceptible difference between the first and the second rating session, $M = 2.44$ to $M = 2.26$. While TA and FLIN both decreased in perceived difficulty, the criterion RSL remained stable (+ 0.01%) and ASL increased only slightly (+ 0.08%) between the two rating sessions.

Table 4.26 Perceived difficulty of applying the four sub-scales

Criterion	Session	Very difficult		Difficult		Easy		Very easy	
		N	Row %	N	Row %	N	Row %	N	Row %
TA	1	8	21.1	17	44.7	13	34.2	0	0.0
	2	4	10.3	15	38.5	19	48.7	1	2.6
	Overall	12	15.6	32	41.6	32	41.6	1	1.3
FLIN	1	1	2.6	15	38.5	23	59.0	0	0.0
	2	0	0.0	11	28.2	27	69.2	1	2.6
	Overall	1	1.3	26	33.3	50	64.1	1	1.3
RSL	1	0	0.0	27	69.2	11	28.2	1	2.6
	2	3	7.7	22	56.4	13	33.3	1	2.6
	Overall	3	3.8	49	62.8	24	30.8	2	2.6
ASL	1	3	7.7	19	48.7	15	38.5	2	5.1
	2	2	5.1	22	56.4	15	38.5	0	0.0
	Overall	5	6.4	41	52.6	30	38.5	2	2.6

Figure 4.6 Comparison of mean perceived difficulty for each criterion and session



4.6.4 Open-ended Justifications

This section will summarize the themes which emerged from the justifications raters provided for their judgment of criterion difficulty (see previous section). The statements were explored for what they might reveal about the rating scale and the

rating process as perceived by the raters, and the features that contributed to greater ease or difficulty of rating certain criteria (see Appendix F for code book).

Overall 1,354 segments were coded via a thematic analysis. The most relevant code family for subsequent analysis were the aspects mentioned to either facilitate ($N = 134$) or inhibit ($N = 179$) rating decisions in relation to the rating scale.

Inhibitors of rating. The factors that seem to limit or inhibit rating from the raters' perspective could be grouped into two major themes. Either participants reported problems with taking a rating decision and expressed doubt or insecurity in their rating decisions (74.9%); or they reported problems related to multi-tasking (22.9 %) (see Table 4.27 for examples and Table 4.28 for frequencies). Almost all segments from the open-ended responses of the survey appeared to fall within one of these two categories. Only a fraction of the segments (Unspecified comments = 2.2 %) reported on difficulties but could not be placed more specifically.

Performance is difficult to rate. This code was used for about a quarter of all segments (26.26%) addressing factors inhibiting rating and distinguished into two sub-codes. The first was used, whenever participants mentioned that performances were uneven regarding a certain dimension (18.44%). This was the case if comments mentioned a discrepancy between strong and weak features (Example 1.1.1) or a fluctuation in quality over the course of the performance. The second code was used when the participants felt there was a tangible mismatch between the descriptors in the scale and the performances (7.8%). This was particularly frequent in those performances that were too short and when the interlocutor intervened and asked the speakers to elaborate. This category was also allocated when raters felt that the performance did not really match the level expressed through the

descriptors in the rating scale (Example 1.1.2). In terms of percentage, mismatch was mentioned slightly less in the second rating session (Session 1 = 9.28% to Session 2 = 6.10%) while the first sub-code remained stable across sessions (Session 1 = 18.56%, Session 2 = 18.29%).

Table 4.27 Example comments illustrating inhibitors of rating

Inhibitors	Example
1. Difficulty with decision-making (no confidence in rating)	
1.1 Performance is difficult to rate	
1.1.1 Uneven performances	Sometimes the lexical range was well developed but it lacked in complexity (Kimberly, S1, RSL)
1.1.2 Mismatch between scale and performance	I had the feeling that many students stay in their comfort zone when speaking and don't try to use more difficult structures. (Jennifer, S2, ASL)
1.2 Awareness of flawed decision-making	
1.2.1 Issues with subjectivity & severity	Rather difficult to be objective due to personal influence: do I like the voice, the tone? (Donny, S2, FLIN)
1.2.2 Issues with differentiating constructs	Moreover, when someone has a very good use of grammar and language or a very natural pronunciation, it is more likely to give a lot of points for all criterion (North, S1, TA)
1.3 Lacking knowledge (right/wrong, level, topic)	
1.3.1 Knowledge of ability levels	but I am not quite sure if their grammar was at a sufficient level or if it was just me who lacks knowledge about the supposed proficiency at level B2. (Jennifer, S1, ASL)
1.3.2 Interpretation of scale	It was difficult to distinguish between a 'wide range' of vocabulary and a 'good' range. (Scarlett, S2, RSL)
2. Difficulty with cognitive demands and attention (no confidence in perception)	
2.1 Identifying salient features	Especially judging whether they used various expressions or good circumlocutions was hard (Lexi, S2, RSL)
2.2 Issues with multi-tasking	Because sometimes I was not sure how much the person said about the different aspects of the task, especially when I was noting something down for RSL or ASL (Esme, S1, TA)
2.3 Complexity of criterion	I still believe that this is the hardest criterion, as it deals with so many different aspects regarding spoken language (Margarete, S2, ASL)

Note. S1 = First rating session, S2 = Second rating session.

Awareness of flawed decision-making. This code consisted of two sub-codes and made up about a fifth of all segments coded as inhibitors (19.55%). Overall, this code was allocated when participants expressed concern about their decision-making with a certain criterion (8.38%) or when they struggled with differentiating constructs (11.17%). Some participants for instance, felt uncomfortable when rating a particular criterion because the process appeared to be more subjective (Example 1.2.1) or because they perceived their own level of severity in that criterion as problematic. Whenever participants noted that they had difficulties with keeping certain criteria separate, the segment was coded with the second sub-code (Example 1.2.2). Overall, this category saw the most distinct shifts between rating sessions. Comments about reliability or severity doubled from 6.19% in Session 1 to 10.98% in Session 2, while comments about difficulties with discerning the different criteria halved from 14.43% to 7.32% in the second rating session.

Lacking rater knowledge. Sub-codes from this category were awarded frequently and made up slightly less than a third of all coded segments for inhibitors of rating (29.05%). This code was generally used whenever participants commented that they lacked certain knowledge they felt was instrumental to fulfilling the rating task. Within this category it was possible to differentiate two sub-codes: a) lacking knowledge of what is sufficient or correct enough (15.64%) and b) lacking knowledge of how to interpret the descriptors or bands in the rating scale (13.41%). The first sub-code includes comments about how raters should judge the quality and quantity of ideas, establish the minimum requirements of passing the exam (Example 1.3.1), or identify typical features for the B2 level. The second sub-category included all comments about difficulties of interpreting descriptors (Example 1.3.2). Overall, the two sub-categories remained stable across the rating

sessions. In terms of percentages, there was a slight increase in statements concerning knowledge of the ability levels (13.40% to 18.29%) and a slight decrease in statements about difficulties with scale interpretation (14.43% to 12.20%).

Difficulties with cognitive demands and attention. In almost a quarter of the comments about factors inhibiting rating (22.9%), participants did not problematize aspects of the rating process concerned with relating the descriptors in the scale to the features in the performance, but instead foregrounded that they struggled with the general demands of the rating task as such. These comments included problems with perceiving salient features or differences in the performances (9.50%, Example 2.1), issues with having to direct their attention or multitask while listening to the performances (10.06%, Example 2.2), and dealing with the complexity of a particular criterion (3.35%). Overall, comments falling into this category remained stable over the rating sessions with a slight decrease of statements concerning issues of identifying salient features (10.3% to 8.5%).

Table 4.28 Number and percentage of comments on inhibiting factors

Inhibitors	Session 1		Session 2		Total	
	N	%	N	%	N	%
1. Difficulty with decision-making (no confidence in rating)						
1.1 Performance is difficult to rate						
1.1.1 Uneven performances	18	18.56	15	18.29	33	18.44
1.1.2 Mismatch between scale and performance	9	9.28	5	6.10	14	7.82
Subtotal	27	27.84	20	24.39	47	26.26
1.2 Awareness of flawed decision-making						
1.2.1 Issues with subjectivity & severity	6	6.19	9	10.98	15	8.38
1.2.2 Issues with differentiating constructs	14	14.43	6	7.32	20	11.17
Subtotal	20	20.62	15	18.29	35	19.55
1.3 Lacking knowledge (right/wrong, level, topic)						
1.3.1 Knowledge of ability levels	13	13.40	15	18.29	28	15.64
1.3.2 Interpretation of scale	14	14.43	10	12.20	24	13.41
Subtotal	27	27.84	25	30.49	52	29.05
2. Difficulty with cognitive demands and attention (no confidence in perception)						
2.1 Identifying salient features	10	10.31	7	8.54	17	9.50
2.2 Issues with multi-tasking	9	9.28	9	10.98	18	10.06
2.3 Complexity of criterion	3	3.09	3	3.66	6	3.35
Subtotal	22	22.68	19	23.17	41	22.91
3. Unspecified	1	1.03	3	3.66	4	2.23
Total	97	100.0	82	100.0	179	100.0

Note. N of documents: n = 39 for Session 1 and 2, N = 78 total.

Facilitators of rating process. The factors that are mentioned as facilitating rating from the participants' perspective were clustered as one code family with several sub-codes (Table 4.28). It was somewhat more difficult to interpret and categorize comments justifying why rating a certain criterion was easy rather than difficult as raters were less explicit or clear in trying to explain their justifications. As a result,

the coding tree describing facilitating factors is less differentiated than the coding tree describing inhibiting factors (see previous section).

Decision feature is easy to perceive. About two thirds of the statements (67.91%) highlighted that decision making for a certain criterion was facilitated because it was easier to perceive critical features. This was noted particularly often about fluency (total 27.61%, Example 1) and other language-related aspects like, for instance, lexical range in the performance (20.15%). Particularly in the first session, participants also commented that they felt it was easy to notice issues with the quality of the content that participants produced (Session 1 = 20.63%, Session 2 = 11.27%). A few participants mentioned a training effect in that perceiving certain features improved between sessions.

Table 4.29 Example comments illustrating facilitators of rating

Facilitators	Example
1. Decision feature is easy to perceive	This was probably the easiest criteria for me because you can quickly decide/hear if someone is fluent or not (North, S1, FLIN)
2. Forming expectations	As I had already seen 15 performances before, I knew which vocabulary I could expect. So, I found it easier to make a decision (Maisie, S2, RSL)
3. Decision feature can be documented well	Ticking each bullet point, when done might have helped me to see whether the task was fulfilled or not. (Lexi, R1, TA)
4. Decision feature is clear cut (correct/incorrect)	It was rather easy to rate the grammar and pronunciation as it is either correct or wrong (Sally, S1, ASL)
5. Descriptors are useful	I still think that fluency depends more on your gut-feeling but the descriptors are somewhat helpful to justify your point of view (Betty, S2, FLIN)
6. Fewer descriptors to deal with	I think the fact that I had to stick to only two descriptors made it easier (Megan, S1, FLIN)
7. Rater knowledge	As I already mentioned, I believe that also amateur raters can assess this criterion. (Maisie, S2, FLIN)
8. Unspecified	I think this was one of the easier things to rate. (Violet, S1, FLIN)

Note. S1 = First rating session, S2 = Second rating session

The remaining 32.09% of segments did not foreground perception and were categorized into seven sub-codes. One group that constituted 8.21% of all statements mentioned that some participants felt that being able to document or note down observations about a performance with regards to a certain criterion made rating this criterion easier (Example 3). A group of statements that saw considerable increase from Session 1 to Session 2 (1.59% to 11.27%) illustrated how having certain expectations towards performances also aided in decision-making (Example 2). All remaining categories (decision features are clear cut, descriptors in scale were useful, criterion was less complex than others, rater knowledge, and unspecified) make up a minor 16.81 percent of the remaining codes (Examples 4 to 8).

Table 4.30 Number and percentage of comments on facilitating factors

Facilitators	Session 1		Session 2		Total	
	N	%	N	%	N	%
1. Decision feature easy to perceive						
1.1 Fluency	16	25.40	21	29.58	37	27.61
1.2 Quality of language	13	20.63	14	19.72	27	20.15
1.3 Quality of content	13	20.63	8	11.27	21	15.67
1.4 Improves with practice	0	0.00	3	4.23	3	2.24
1.5 General	1	1.59	2	2.82	3	2.24
Sub-total	43	68.25	48	67.61	91	67.91
2. Forming expectations	1	1.59	8	11.27	9	6.72
3. Decision feature can be documented well	6	9.52	5	7.04	11	8.21
4. Decision feature is clear cut (correct/ incorrect)	3	4.76	2	2.82	5	3.73
5. Descriptors are useful	1	1.59	3	4.23	4	2.99
6. Fewer descriptors to deal with	1	1.59	2	2.82	3	2.24
7. Rater knowledge	1	1.59	1	1.41	2	1.49
8. Unspecified	7	11.11	2	2.82	9	6.72
SUM	63	100.0	71	100.0	134	100.0

Note. N = 39 of documents.

4.6.5 Comparison across criteria

The previous sections described the two main code families, facilitators and inhibitors, which emerged from the thematic analysis of raters' justifications. This section will look more specifically at how the codes from both groups are distributed across the criteria (see Table 4.31 and Table 4.32). In doing so, the analysis focuses on exploring the reasons why certain criteria might appear more difficult or easy to rate.

Task achievement. While raters did not perceive TA as necessarily the most difficult criterion, they associated a range of challenges with rating this criterion. The two main issues appear to be that raters felt that there is some kind of mismatch between scale and performance (78.57%). Compared to the other three criteria, raters were also more concerned with subjectivity when rating this criterion (50.00%). On the other hand, raters had fewer problems with differentiating the construct (8.00%) and lack of certain knowledge in comparison to the other three criteria (knowledge of requirements 25.71%, interpretation of scale 20.00%).

As far as the second large code family was concerned, raters consistently reported issues with directing their attention for the criterion TA (identifying salient features 38.10%, issues with multi-tasking 38.10%). However, raters did not find that this was due to the complexity of the criterion (0.00%). Rather, raters reported struggling with focusing on this criterion (“It is difficult to focus on this”, Amy, S2) or dividing their attention between the criteria (“It’s hard to decide whether to focus more on structure or on content”, Stormi, S2).

In terms of facilitators, many raters indicated that rating TA improved between the two sessions because they had formed a clearer expectation of what speakers will and can do with the tasks (50.00%) and that it was an advantage that one can document the decision feature for this criterion more easily (44.44%; “with checking the bullet points it was quite easy to see whether the students fulfilled all tasks or not”, Lexi, S2).

Fluency and interaction. Raters found this criterion easiest to use and accordingly, there were fewer justifications associated with difficulties of applying the FLIN scale. Raters noted more frequently with this criterion that it was challenging to

differentiate the construct (36%) (Excerpt 1). Within the larger code category 2 concerned with cognitive load there were only a few mentions for identifying salient features (2.1) but no statements for the other two codes (2.2 or 2.3).

Excerpt 1: it was hard to rate those students who were fluent in speech but made a lot of mistakes because do I only rate the fluency or does it get worse when there are a lot of mistakes? (Sally, S1)

In terms of facilitators, most coded segments were elicited in the context of this criterion (total 34.67%, see Table 4.32). Forty percent of all coded segments focusing on ease of perception were mentioned in relation to FLIN. The main reason why raters seem to find it easier to rate this criterion might be that fluency is a feature that can be perceived and possibly judged quickly (Excerpt 2).

Excerpt 2: This was probably the easiest criteria for me because you can quickly decide/hear if someone is fluent or not. (North, S1)

Range of spoken language. Overall, most coded segments describing difficulties pertained to this criterion (30.04%). The most frequent codes were the problem of uneven performances (31.82%), differentiating the construct (28.00%), and lacking knowledge (knowledge of requirements 42.86% and interpretation of the scale 31.43%). Raters struggled with weighing the different descriptors if they detected features that were described in different bands of the scale (Excerpt 3). Comments indicate that a lack of knowing which features are typical of the B2 level (Excerpt 4) and problems of understanding the scale descriptors (Excerpt 5) were particular challenges. On top of this, raters also admitted to struggling with identifying the salient features (38.10%).

Excerpt 3: I find it rather difficult since some of the participants used a wide range of vocabulary but sometimes were not able to use it in the correct context. (Megan, S1)

Excerpt 4: I'm still not sure what can be expected from learners at level B2 what makes it hard to rate that aspect. (Jennifer, S2)

Excerpt 5: ... but what makes is hard is to decide if it's, for example, band 8, 9 or 10, as there is hardly any difference in the wording. (Deborah, S 2)

What seemed to help raters with using this criterion was that noticing the decision criterion improved with practice (50.00%), that it can be documented well (55.56%), and that raters can build up more specific expectations (28.57%) as they gain more experience.

Accuracy of spoken language. Similar to RSL, a big challenge of using this criterion was linked to uneven performances and weighing descriptors (31.82%). Raters were also concerned about their severity with this criterion (25.00%) and their lack of knowledge as far as the requirements (31.43%) and interpretation of the descriptors (28.57%) were concerned. While they struggled less with identifying the salient features for their decision-making (9.52%) and many felt that it was easy to notice these characteristics (Table 4.32, 65.63%), issues with multi-tasking (38.10%) and the complexity of the criterion (62.50%) were mentioned more frequently than with other criteria.

Table 4.31 Distribution of aspects inhibiting rating decisions across criteria (percentage row-wise)

	TA		FLIN		RSL		ASL		Total	
	N	%	N	%	N	%	N	%	N	%
1 Difficulty with decision making	43	25.44	29	17.16	50	29.59	47	27.81	169	100.00
1.1 Performance is difficult to rate	17	29.31	11	18.97	15	25.86	15	25.86	58	100.00
1.1.1 Uneven performances	6	13.64	10	22.73	14	31.82	14	31.82	44	100.00
1.1.2 Mismatch between scale and performance	11	78.57	1	7.14	1	7.14	1	7.14	14	100.00
1.2 Awareness of flawed decision-making	10	24.39	11	26.83	9	21.95	11	26.83	41	100.00
1.2.1 Issues with subjectivity & severity	8	50.00	2	12.50	2	12.50	4	25.00	16	100.00
1.2.2 Issues with differentiating constructs	2	8.00	9	36.00	7	28.00	7	28.00	25	100.00
1.3. Lacking knowledge (right/wrong, level, topic)	16	22.86	7	10.00	26	37.14	21	30.00	70	100.00
1.3.1 Knowledge of requirements	9	25.71	0	0.00	15	42.86	11	31.43	35	100.00
1.3.2 Interpretation of descriptors	7	20.00	7	20.00	11	31.43	10	28.57	35	100.00
2 Difficulty with directing attention and noticing	16	32.00	3	6.00	16	32.00	15	30.00	50	100.00
2.1 Identifying salient features	8	38.10	3	14.29	8	38.10	2	9.52	21	100.00
2.2 Issues with multi-tasking	8	38.10	0	0.00	5	23.81	8	38.10	21	100.00
2.3 Complexity of criterion	0	0.00	0	0.00	3	37.50	5	62.50	8	100.00
Unspecified	1	25.00	1	25.00	1	25.00	1	25.00	4	100.00
TOTAL	60	26.91	33	14.80	67	30.04	63	28.25	223	100.00

Table 4.32 Distribution of aspects facilitating rating decisions across criteria (percentage row-wise)

	TA		FLIN		RSL		ASL		Total	
	N	%	N	%	N	%	N	%	N	%
1. Decision feature easy to perceive	23	23.00	40	40.00	14	14.00	23	23.00	100	100.00
1.1 Fluency	0	0.00	38	100.00	0	0.00	0	0.00	38	100.00
1.2 Quality of language	0	0.00	0	0.00	11	34.38	21	65.63	32	100.00
1.3 Quality of content	22	100.00	0	0.00	0	0.00	0	0.00	22	100.00
1.4 Improves with practice	1	25.00	0	0.00	2	50.00	1	25.00	4	100.00
1.5 General	0	0.00	2	50.00	1	25.00	1	25.00	4	100.00
2. Decision feature can be documented well	4	44.44	0	0.00	5	55.56	0	0.00	9	100.00
3. Forming expectations	7	50.00	2	14.29	4	28.57	1	7.14	14	100.00
4. Decision feature is clear cut (correct/ incorrect)	1	16.67	0	0.00	1	16.67	4	66.67	6	100.00
5. Descriptors are useful	2	40.00	2	40.00	0	0.00	1	20.00	5	100.00
6. Criterion less complex	0	0.00	3	100.00	0	0.00	0	0.00	3	100.00
7. Rater knowledge	1	33.33	2	66.67	0	0.00	0	0.00	3	100.00
8. Unspecified	0	0.00	3	30.00	3	30.00	4	40.00	10	100.00
TOTAL	38	25.33	52	34.67	27	18.00	33	22.00	150	100.00

4.6.6 Summary Perception Data

The perception data was collected to explore how the novice raters felt about applying the rating scale and how confident they were about their rating decisions.

General confidence. The raters self-reported a moderate degree of confidence in their rating decisions overall. There was a trend of growing confidence as the sessions progressed. This is evidenced by more than 75% of the group reporting that they felt somewhat confident (44.24%) or moderately confident (32.10%) in their rating.

General perceived difficulty of rating. There was a steady improvement in terms of general perceived difficulty of rating the performances throughout the experiment. On average, around half of the responses reported rating to be slightly difficult (47.44%). However, at the end of the experiment more than 60% (23 raters) were neutral (neither easy nor difficult) or even reported finding rating to be slightly easy.

Difficulty of rating criteria in the scale. Overall, raters found the criterion TA the most difficult and FLIN the easiest to rate, with the criteria RSL and ASL taking a middle position between the other two dimensions of the scale. There was, however, a quite dramatic shift concerning TA between rating sessions, where it moved from 1st position to 3rd.

The analysis of justifications that raters gave to support their judgement of criterion difficulty provided further details about how the novice raters used the rating scale. What made the criterion TA difficult to rate was a frequently perceived mismatch between performances and rating scale as well as raters noticing issues with their

own levels of severity and subjectivity. Justifications for difficulties with RSL and ASL frequently foregrounded uneven performances and a perceived lack of knowledge about what features would constitute the expected target level. The criteria ASL and RSL were also found to be particularly complex. Finally, the comments revealed that dividing and directing attention was an issue when rating TA and ASL. Regarding the criterion FLIN, many raters mentioned that the construct was less concrete than for other criteria. However, fluency can be observed easily in the performances and the scale included fewer descriptors.

4.7 Summary

This chapter reported on the results of Study 1, which focussed on gathering data to support the evaluation inference and investigated how a group of novice raters used the Austrian rating scale when assessing speaking performances. The quantitative analyses showed great inter-rater variation as far as the raters' severity and accuracy were concerned. However, even though the raters were novices with little training and hardly any classroom experience, about half of the group ($n = 19$) obtained moderate consensus in terms of kappa and moderate ($n = 20$) to strong ($n = 16$) consistency with the fair scores. The qualitative analysis revealed that raters developed greater confidence and ease as the experiment progressed. Some of the challenges raters appeared to face concerned their perceived lack of knowledge of the target level, matching the performances to the scale, dividing their attention and self-awareness. An important finding across various analyses was the differential functioning of the TA criterion which was also confirmed by the raters' perception of this criterion. Raters also felt that rating FLIN was easier which was confirmed by higher accuracy on this criterion. Interestingly, the criteria RSL and ASL, which

208

raters' continued to perceive as difficult, emerged as easier to rate accurately than the criterion TA.

5 Results II: Exploring the Role of Cognitive and Psychological Attributes in Rater Cognition

5.1 Outline

The second study explored whether traits such as cognitive attributes, cognitive processing preferences or decision-making style might affect rater cognition. Rater cognition was captured through rating quality (i.e., rater severity and accuracy) and rater behaviour metrics (i.e., process data based on time stamps: deliberation time, time to first decision and number of revisions). The analyses mainly involved correlating dependent variables (i.e., rater cognition variables) with the independent variables (see variable map in Appendix A).

In the following sections, the results will be presented according to each independent variable group. First, section 5.2 reports the descriptive statistics for the time stamp data. Next, section 5.3 reports the descriptive data from the five cognitive tests and how they correlated with measures of rater quality and rater behaviour. The final two sections, 5.4 and 5.5, present the results for the two psychological questionnaires (REI-40 and GDMSI) and the correlation of raters' REI-40 and GDMSI scores with rating quality and rater behaviour metrics. The chapter ends with a summary of the key results.

5.2 Rating behaviour metrics

5.2.1 Deliberation Time (DT)

For the purposes of this study, deliberation time was defined as the time a rater spent on rating a performance. The variable was based on the total viewing time (i.e., timestamp of starting to view a performance to timestamp of submitting all four rating decisions) minus the length of the performance to correct for differences in performance length. The descriptive statistics for the mean deliberation times are summarized in Table 5.1.

Table 5.1 Descriptive statistics for mean deliberation time (measured in seconds) for each set of performances and session

Variables	Min	Max	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
Set 1	7.8	315.27	91.10	71.48	1.15	1.13
Set 2	-2.67	374.39	55.48	62.30	3.72	18.14
Session 1	2.56	344.83	73.29	62.50	2.38	8.40
Set 3	5.33	177.39	53.66	46.17	1.44	1.35
Set 4	3.28	158.20	39.21	37.62	1.73	2.77
Session 2	4.30	150.03	46.43	38.18	1.26	0.97
Total	5.23	207.97	59.86	45.68	1.49	2.08

Note. *SE for Skewness* = 0.38; *SE for Kurtosis* = 0.74.

Each of the two rating sessions presented the raters with 15 performances and after the first eight speakers within each session (classified as a “Set”), participants were asked about the rating process and reminded to take a break if needed. Therefore, deliberation times can be summarized meaningfully in several ways: (1) after each set of performances, providing a more dynamic snapshot, (2) after each rating session, or (3) with an overall mean.

Intercorrelations between the individual sets (x 4) and overall sessions (x 2) were inspected to decide which variables should be used for later analyses. Spearman rank correlations between the variables within the same session were strong and significant between Set 1 and 2 ($r_s = .661, p < .001$) and between Set 3 and 4 ($r_s = .836, p < .001$). Correlations between sets and overall session mean times were even more pronounced. Set 1 and 2 had correlation coefficients of $r_s = .951 (p < .001)$ and $r_s = .834 (p < .001)$, respectively with the overall mean deliberation time of Session 1. Mean times for Set 3 and 4 had correlation coefficients of $r_s = .975 (p < .001)$ and $r_s = .917 (p < .001)$ with the overall mean deliberation time of Session 2 (intercorrelation matrix in Appendix G). In light of these results, a composite variable constituting the overall mean deliberation time across both sessions was used for further procedures.

5.2.2 Time to First Decision (TTFD)

This variable was defined as duration between the timestamp capturing the start of rating a performance and the timestamp of entering the first decision by selecting a box in the online rating grid representing the rating scale. See Table 5.2 for a summary of the descriptive statistics. There was considerable variation in how much time passed before raters selected their first decision in the rating grid which is evident in the great range between the minimum and maximum values. However, overall, there is less variation visible in the mean values between the sessions than in the data for deliberation time. There is a slight trend of data points becoming less dispersed in later rating sessions indicated by a decreasing *SD*. On average for Session 2, raters took their first decision about four minutes into the performance ($M_{TTFD} = 254.42$ seconds). Correlating the mean times of each set with the overall

session variable produced strong coefficients ($r_s = .790$, $p < .001$, or higher; intercorrelation matrix in Appendix G). A composite variable was created by averaging the mean times to first decision for both sessions.

Table 5.2 Descriptive statistics for mean time to first decision (measured in seconds) for each set of performances and each session

Variables	Min	Max	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
Set 1	49.95	452.09	289.46	100.69	-0.70	-0.05
Set 2	37.30	604.19	250.88	104.10	0.60	2.22
Session 1	43.62	487.69	270.17	96.08	-0.35	0.03
Set 3	47.35	418.82	266.15	88.41	-0.51	-0.41
Set 4	41.94	418.23	242.69	91.39	-0.35	-0.65
Session 2	44.65	418.53	254.42	87.34	-0.43	-0.40
Total	44.13	418.57	262.30	89.02	-0.44	-0.20

Note. *SE for Skewness = 0.38; SE for Kurtosis = 0.74.*

5.2.3 Revision Count (RC)

Revisions were identified by looking at the number of clicks in the online rating form. As the rating scale has four criteria, the minimum number required to submit the rating decisions was four mouse clicks. Any additional click that was recorded by the survey was categorized as revision. Similar to data for time to first decision (previous section), there were substantial differences in the extent to which raters used the opportunity to revise their ratings (see Table 5.3). While some raters never changed any of their decisions, others changed them up to 21 times – which would amount to an average of at least twice per performance – during a rating session (see RSL Session 1). Most subsequent analyses were based on an overall mean count across both sessions. However, the criterion-specific variables were used whenever criterion-specific severity and accuracy were investigated as the number of revisions may be related to rater accuracy or severity.

Table 5.3 Descriptive statistics for revision count across sessions

Session	Variables	Min	Max	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
1	TA	0	17	3.23	4.08	1.94	3.71
	FLIN	0	16	3.41	4.45	1.74	2.49
	RSL	0	21	3.46	4.51	2.13	5.36
	ASL	0	17	3	3.84	2.03	4.37
	Overall	0	64	13.1	15.31	1.87	3.23
2	TA	0	13	3.54	3.78	0.95	-0.23
	FLIN	0	14	3.28	4.16	1.46	1.2
	RSL	0	18	3.46	4.37	1.92	3.58
	ASL	0	18	3.18	4.62	1.67	2.21
	Overall	0	59	13.46	14.87	1.44	1.89
Total		0	123	26.46	28.96	1.76	3.14

Note. *SE* for Skewness = 0.38; *SE* for Kurtosis = 0.74.

5.2.4 Intercorrelations Between Dependent Variables

Severity. A correlational analysis revealed there was no significant relationship between variables of rater severity and variables based on rater behaviour metrics. For deliberation time, Spearman rank correlations ranged from $r_s = -.09$ (severity ASL) to $r_s = -.28$ (severity FLIN). For time to first decision, coefficients ranged from $r_s = -.01$ (severity full model) to $r_s = -.13$ (severity FLIN). Finally, for revisions, the coefficients were between $r_s = -.07$ (severity FLIN) to $r_s = -.25$ (severity TA). Based on these results there was not enough evidence to dismiss the null hypotheses that severity overall or on any specific criterion might be associated with any of the rater behaviour metrics (intercorrelations in Appendix H).

Accuracy. The Spearman rank analyses revealed moderate and statistically significant association between rater accuracy and one of the rater behaviour metrics. Rater accuracy on the criterion TA had a moderately strong association with the number of revisions ($r_s = .56, p < .001$). Thus, a higher number of revisions coincided with higher accuracy in the criterion TA.

The remaining coefficients remained weak to very weak. For deliberation time, values for *rho* ranged from $r_s = -.01$ (accuracy TA) to $r_s = .21$ (accuracy FLIN). Disregarding the significant correlation with accuracy TA, correlations for time to first decision and accuracy were between $r_s = .03$ (accuracy FLIN) and $r_s = -.44$ (accuracy TA), and for revisions and accuracy between $r_s = .17$ (accuracy RSL) and $r_s = .39$ (accuracy full model). Of all three rater behaviour metrics, the number of revisions appeared to be the most promising predictor of rater accuracy (intercorrelations in Appendix H).

5.3 Cognitive Attributes

5.3.1 Descriptive Statistics

The preliminary exploration of the variables (see also Analysis 3.8.6) showed approximately normal distributions for the Stroop, Letters-numbers and Keep Track task. The Shapiro-Wilk test indicated non-normal distributions for the Digit Span (.898, $p = .002$) and Trail Making tasks (.856, $p < .001$). There were two extreme observations. Amy performed extraordinarily strong in the Stroop task and Sally scored extremely poorly in the Trail Making task (see Table 5.4 for descriptive statistics).

Compared to other data available on the cognitive tests, the performance of this cohort appears typical. They performed stronger in the Digit Span and slightly weaker in the Keep Track tasks than the 19–21-year-old student sample in Indrarathne and Kormos (2017), $M_{Digit\ Span} = 5.05-5.70$ ($SD = 1.35-1.72$) and $M_{KeepTrack} = 70.21-77.40$ ($SD = 13.53-17.26$). The mean Stroop effect was between 1822.15 to 2150 ms in Indrarathne and Kormos' (2017) groups which is a lot more

than the mean effect prior to reversing the variable in this cohort ($M = 90.26$, $SD = 89.46$) or Miyake et al. (2000), $M = 166$ ms, $SD = 60$. As both, the current study as well as Indrarathne and Kormos (2017) relied on the output from different websites there may have been differences in how the effects were calculated, displayed, or transformed. Other reasons for this discrepancy could be differences in keyboarding skills or that this study provided the Stroop in the participants' first language German rather than English. This cohort's performance on the Trail Making tasks A and B ($M_{Trial A} = 24.06$, $SD = 4.78$; $M_{Trial B} = 52.15$, $SD = 17.21$) compares well to the normative data for 25-34-year-olds collated by Tombaugh (2004), $M_{Trial A} = 24.40$, $SD = 8.71$; $M_{Trial B} = 50.68$, $SD = 12.36$). Thus, the mean difference between trials in this cohort ($M = 26.34$, $SD = 12.03$) can be considered typical.

Table 5.4 Descriptive statistics for cognitive tasks (N = 39)

Variables	Min*	Max	M	SD	Skewness	Kurtosis
Stroop	0	376	159.18	88.34	0.32	0.04
Letters-numbers	0	810	471.95	204.71	-0.17	-0.93
Digit Span	5	9	7.10	1.05	-0.36	-0.54
Keep Track	37.04	96.30	67.24	12.59	-0.11	0.01
Trail Making	0	90.59	66.80	16.17	-1.90	6.44

Note. * Stroop, Letters-numbers and Trail Making task were reversed to align direction of all cognitive variables. SE for Skewness = 0.38; SE for Kurtosis = 0.74.

Each cognitive test was expected to target a separate component of executive functioning and attention. A Spearman rank correlational analysis was carried out to confirm this assumption (see Table 5.5). There were no significant correlations between any of the cognitive variables. In many cases there was no discernible association between the variables at all; for example, between the Stroop and the Keep Track task. There was one weak positive association between the Trail Making and the Keep Track task ($r_s = .32$). As each cognitive variable was intended

to measure a somewhat distinct component of executive functioning and attention, this was a desirable outcome.

Table 5.5 Intercorrelations between cognitive test scores

Variable	1	2	3	4
1. Stroop				
2. Letters-Numbers	.27 [-.07, .56]			
3. Digit Span	-.25 [-.53, .06]	-.10 [-.27, .41]		
4. Keep Track	-.05 [-.35, .26]	.08 [-.27, .41]	-.13 [-.42, .19]	
5. Trail Making	-.29 [-.61, .06]	-.03 [-.43, .23]	.06 [-.23, .37]	.32 [.03, .57]

Note. Values in square brackets indicate the 95% confidence interval for each correlation.

5.3.2 Cognitive Attributes and Rating Quality

Each aspect of rating quality (i.e., severity, fit and accuracy) was correlated with each cognitive variable to explore any noteworthy associations in the data.

Severity. There were no statistically significant relationships between cognitive variables and measures of rater severity (see Table 5.6). Most correlation coefficients were very small, ranging from $r_s = .01$ ($p = 1$, Letters-Numbers x Severity FLIN) which was one of the weakest association to $r_s = -.25$ ($p = 1$, Trail Making x Severity RSL) which was the strongest association. The Stroop task overall displayed very weak and quite similar negative associations with all measures of severity. The correlation coefficients for the Letters-Numbers task with severity range from virtually unrelated (e.g., full model $r_s = .01$, $p = 1$) to very weakly associated (RSL $r_s = -.15$, $p = 1$). The Digit Span task had no discernible association with variables of rater severity, with correlation coefficients ranging

from $r_s = -.02$ to $r_s = -.06$. The associations between the Keep Track task and rater severity show a clear but weak negative pattern, $r_s = -.17$ with TA to $r_s = -.20$ with RSL. Finally, there was no discernible relationship between the Trail Making task and TA or FLIN, but the association with severity on the two language related criteria RSL and ASL emerged as more pronounced, $r = -.25$ ($p = 1$) and $r = .21$ ($p = 1$) respectively.

Table 5.6 Spearman correlations between rater severity and cognitive tasks

Variable	Stroop	Letters-Numbers	Digit Span	Keep Track	Trail Making
Severity (Full model)	-.20 [-.50, .16]	-.01 [-.38, .36]	-.05 [-.32, .23]	-.17 [-.50, .16]	-.15 [-.48, .19]
Severity (TA)	-.18 [-.51, .18]	.06 [-.34, .43]	-.03 [-.32, .29]	-.16 [-.42, .14]	-.01 [-.31, .31]
Severity (FLIN)	-.20 [-.50, .11]	.01 [-.33, .34]	-.04 [-.33, .26]	-.13 [-.44, .19]	-.08 [-.42, .26]
Severity (RSL)	-.22 [-.50, .10]	-.15 [-.47, .17]	-.02 [-.32, .25]	-.20 [-.54, .16]	-.25 [-.55, .10]
Severity (ASL)	-.20 [-.47, .11]	-.06 [-.37, .25]	-.06 [-.35, .25]	-.15 [-.48, .21]	-.21 [-.48, .10]

Note. Values in square brackets indicate the 95% confidence interval for each correlation.

The results from this analysis make it appear unlikely that any of the components measured by the cognitive tests had a strong impact on rater severity as operationalized in this study. Out of all five, the Keep Track and Stroop tasks displayed the strongest correlation coefficients overall. It is noteworthy that most associations were negative in that a lower score in the cognitive tasks coincided with a higher level of severity. However, none of the relationships explored were statistically significant.

Accuracy. As was the case for the analysis based on the severity measures, there were no statistically significant relationships between the five cognitive variables and MFRM estimates of rater accuracy (see Table 5.7). Many coefficients were remarkably weak. In relative terms, the Stroop test produced the strongest and most consistently positive coefficients of all the cognitive variables and the overall highest association with accuracy in FLIN, $r_s = .27$ ($p = .297$). Performing stronger on this task appeared to be associated with higher rating accuracy. In the Letters-Numbers task, the majority of the coefficients were positive and either very weak (TA, FLIN, RSL) or there was no association at all (full model, ASL). As was the case for severity, the Digit Span task overall produced the lowest associations with measures of rater accuracy and appears unrelated to rating quality. The Keep Track task mainly produced positive correlation coefficients, particularly with the accuracy on ASL ($r_s = .21$, $p = .670$). The Trail Making Task appeared unrelated to accuracy of rating TA, FLIN or the full model, but correlated positively with accuracy on the two language-related criteria RSL and ASL. Finally, it is noteworthy that the accuracy (TA) variable produced distinctly different correlation coefficients for both, the Letters-Numbers task and the Keep Track task.

Correlation patterns between cognitive variables and rating accuracy are similar to the patterns with rater severity as the coefficients were mainly weak or very weak for both analyses.

Table 5.7 Spearman correlations between rater accuracy and cognitive tasks

Variable	Stroop	Letters-Numbers	Digit Span	Keep Track	Trail Making
Accuracy (Full model)	.23 [-.10, .52]	.06 [-.28, .40]	-.01 [-.28, .27]	.03 [-.32, .35]	.06 [-.22, .34]
Accuracy (TA)	.14 [-.19, .45]	-.13 [-.45, .22]	-.11 [-.44, .20]	-.17 [-.48, .17]	-.05 [-.35, .28]
Accuracy (FLIN)	.27 [-.02, .52]	.18 [-.18, .51]	-.02 [-.33, .27]	.04 [-.28, .38]	-.04 [-.35, .26]
Accuracy (RSL)	.23 [-.11, .54]	.17 [-.21, .53]	-.03 [-.30, .26]	.11 [-.26, .45]	.19 [-.13, .50]
Accuracy (ASL)	.15 [-.24, .47]	.07 [-.28, .43]	.07 [-.19, .32]	.21 [-.11, .47]	.22 [-.10, .48]

Note. Values in square brackets indicate the 95% confidence interval for each correlation.

5.3.3 Cognitive Attributes and Rating Behaviour

Deliberation time. There were no significant correlations between cognitive measures and deliberation time. Table 5.8 below presents the correlation coefficients which range from $r_s = .02$ with the Digit Span task to $r_s = .37$ ($p = .24$) with the Stroop. The Digit Span task appears unrelated to deliberation time with a particularly small correlation coefficient close to 0. The correlation coefficients between deliberation time and the Stroop task scores show a weak positive relationship. The scores on the Letters-Numbers task as well as the Trail Making task, which tap into task switching and attention focus, show a weak negative association with deliberation time.

Table 5.8 Spearman correlations between deliberation time (DT) and cognitive tasks

Variable	Stroop	Letters-Numbers	Digit Span	Keep Track	Trail Making
<i>M</i> DT	.37	-.25	.02	.11	-.29
	[.07, .61]	[-.58, .01]	[-.32, .36]	[-.23, .43]	[-.59, .05]

Note. Values in square brackets indicate the 95% confidence interval for each correlation.

Time to first decision (TTFD). According to Table 5.9, there are no significant associations between the mean time to first decision and measures from the cognitive tests. Even the most pronounced correlation coefficients, which range from $r_s = -.19$ with the Letters-Numbers task to $r_s = .21$ with the Keep Track task, are negligible. Overall, the predictive power of the cognitive tasks to when raters take down their first decision was very weak. None of the reported associations were statistically significant.

Table 5.9 Spearman correlations between time to first decision (TTFD) and cognitive tasks

Variable	Stroop	Letters-Numbers	Digit Span	Keep Track	Trail Making
<i>M</i> TTFD	.18	-.19	-.06	.21	-.18
	[-.15, .44]	[-.52, .11]	[-.37, .30]	[-.10, .51]	[-.47, .16]

Note. Values in square brackets indicate the 95% confidence interval for each correlation.

Number of revisions. The associations between the cognitive measures and the total number of revisions are presented in Table 5.10. The correlation coefficients were generally weak to negligible; the smallest coefficient was $r_s = -.01$ with the Digit Span and the strongest coefficient was $r_s = -.37$ with the Keep Track task. The Letters-Numbers task, Digit Span task, Trail Making task and Stroop task all show very weak correlation coefficients and appear unrelated to the number of revisions.

Table 5.10 Spearman correlations between number of revisions and cognitive tasks

Variable	Stroop	Letters-Numbers	Digit Span	Keep Track	Trail Making
Revisions	0.13	-.05	-.01	-.37	-.10
	[-.23, .47]	[-.39, .27]	[-.36, .35]	[-.63, -.07]	[-.45, .25]

Note. Values in square brackets indicate the 95% confidence interval for each correlation.

5.4 Preferred Cognitive Processing Mode (REI-40)

5.4.1 Descriptive Statistics and Scale Properties

Table 5.11 summarizes the results from the REI-40 questionnaire. Overall, participants expressed a slight preference for experiential processing ($M_{experiential} = 3.73$, $SD = .54$) compared to rational processing ($M_{rational} = 3.66$, $SD = .41$). The rational engagement (RE) subscale achieved the highest mean ($M_{RE} = 3.77$, $SD = .47$). Rational ability (RA) was the least preferred sub-scale ($M_{RA} = 3.55$, $SD = .47$).

Table 5.11 Descriptive statistics for Rational Experiential Inventory (REI-40)

Variables	Min	Max	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
Rational Ability (k = 10)	2.50	4.80	3.55	0.47	0.09	0.19
Rational Engagement (k = 10)	2.60	4.60	3.77	0.47	-0.10	-0.29
Rationality (k = 20)	2.80	4.65	3.66	0.41	0.15	-0.09
Experiential Ability (k = 10)	2.60	4.90	3.74	0.61	-0.10	-0.63
Experiential Engagement (k = 10)	2.80	4.80	3.72	0.53	-0.18	-0.64
Experientiality (k = 20)	2.75	4.80	3.73	0.54	-0.13	-0.77

Note. *SE* for Skewness = 0.38; *SE* for Kurtosis = 0.74.

The total scale reliability coefficients between the two processing modes were comparable (Rationality, $\alpha = .80$, Experientiality, $\alpha = .89$). There were no significant correlations between the Rationality and Experientiality scales (see

Table 5.12), which is in line with Pacini and Epstein’s data (1999). This confirms the tenet of dual-mode processing theory that the two information processing modes are independent from each other. There was a moderate correlation between rational engagement and ability, but strong correlations between all subscales with the overall main scales. This justifies using the combined Experientiality and Rationality scales as well as retaining the subscales for Rational Ability and Engagement. Given the rather high correlation between the two Experiential subscales ($r = .82, p < .01$), however, there is little empirical support in this dataset to consider these constructs as clearly distinct. In light of these results only the mean scores on the two overall scales (i.e., Rationality and Experientiality) were retained for subsequent correlational analyses.

Table 5.12 Intercorrelations and reliability coefficients for Rational Experiential Inventory (REI-40)

	1	2	3	4	5	6
Rational Ability	(.74)					
Rational Engagement	.46**	(.67)				
Rationality (overall)	.85**	.85**	(.80)			
Experiential Ability	.14	.07	.11	(.84)		
Experiential Engagement	.09	-.01	.01	.82**	(.78)	
Experientiality (overall)	.12	.03	.06	.96**	.94**	(.89)

Note. $N = 39$. Reliabilities for the REI scales (Cronbach’s Alpha) in parentheses. ** Correlation is significant at the 0.01 level (2-tailed).

In addition, the Processing Style Influence (PSI) score (Gunnell & Ceci, 2010; see also Analysis 3.8.5) was calculated for all participants. The formula as suggested by Gunnell and Ceci (2010) was

$$\text{PSI} = [(Mdn \text{ cohort rationality}) - (\text{individual rationality score})] + [(\text{individual experientiality score}) - (Mdn \text{ cohort experientiality})]$$

Deborah, for example, had an individual rationality score of 65 and experientiality score of 89. The cohort medians were 73 for rationality and 76 for experientiality. Plugging these values into the PSI formula gives Deborah a PSI score of 21: $[(73-65) + (89-76)]$.

Table 5.13 Rational and experiential processors based on PSI scores

Rational Processors		Experiential Processors	
Name	PSI score	Name	PSI score
North	-28	Deborah	21
Scarlett	-28	Maisie	18
Esme	-20	Dorothy	18
Hendrix	-20	Donny	15
Margaret	-16	Alice	15
Rosie	-15	Violet	13
Zara	-14	Lexi	13
Helen	-12	Tyler	11
Layla	-12	Holly	10
Paige	-11	Kimberly	9
Mary	-10	Sally	7
Megan	-10	Willow	6
Susan	-9	Daisy	4
Phoebe	-8	Stormi	3
Amber	-6	Eliza	2
Zachary	-6	Penelope	1
Jennifer	-6	Betty	1
Lucy	-5	Maddison	0
Nancy	-2		
Amy	-2		
Chloe	-2		

Note. PSI = Processing Style Influence.

The PSI scores as calculated for all 39 participants are presented in Table 5.13. The participants with the strongest orientation towards a given processing style and a higher PSI are listed toward the top. With high negative scores, raters such as North,

Scarlett, Esme and Hendrix emerge as strongly oriented towards rational processing while raters like Deborah, Maisie and Dorothy were found to orient towards experiential processing.

5.4.2 Preferred Cognitive Processing Mode and Rating Quality

The data from the REI-40 questionnaire produced continuous as well as categorical variables. The scores for each rater on the Experiential and Rational scales were then correlated with measures of rating quality. The PSI scores were used as the independent, categorical variable in non-parametric group comparisons.

Severity. The correlation matrix for rater severity with the Rationality and Experientiality scores is provided in Table 5.14. There were no significant associations. Overall, all correlation coefficients are below $\pm .2$, and can be considered non-existent. Relatively speaking, the highest positive correlation coefficient was found between severity (TA) and rationality ($r_s = .15, p = 1$). The lowest negative correlation was between severity (RSL) and experientiality ($r_s = -.13, p = 1$). Based on these findings, preferred processing mode could not be established to be a predictor of rater severity.

A Mann-Whitney U test was carried out to investigate whether any differences between the two modes of processing were statistically significant using the categorical variable based on the PSI transformation (see Table 5.15 for summary). Distributions between the two groups appeared similar upon visual inspection. There were no significant differences in severity levels between the two processing groups. The null hypothesis that both groups (rational processors and experiential processors) rate similarly in terms of severity was therefore retained.

Table 5.14 Spearman correlations between severity and preferred cognitive processing mode (N = 39)

Variable	Rationality	Experientiality
Severity (full model)	.00 [-.36, .27]	-.09 [-.40, .23]
Severity (TA)	.15 [-.18, .45]	-.08 [-.37, .21]
Severity (FLIN)	-.07 [-.38, .24]	-.01 [-.31, .29]
Severity (RSL)	-.13 [-.45, .20]	-.15 [-.46, .18]
Severity (ASL)	-.06 [-.39, .27]	-.09 [-.40, .23]

Note. TA = Task achievement, FLIN = Fluency, RSL = Range of spoken language, ASL = Accuracy of spoken language. Values in square brackets indicate the 95% confidence interval for each correlation.

Table 5.15 Summary of Man-Whitney U tests between rational and experiential processors and rating severity

		Rational processors (N = 21)	Experiential processors (N = 18)	p-value
Severity (full model)	Mdn.	1.57	1.54	.707
	Avg. rank	20.64	19.25	
Severity TA	Mdn.	1.54	1.45	.791
	Avg. rank	20.45	19.47	
Severity FLIN	Mdn.	1.96	1.76	.856
	Avg. rank	20.33	19.61	
Severity RSL	Mdn.	1.57	1.51	.686
	Avg. rank	20.71	19.17	
Severity ASL	Mdn.	1.46	1.74	.666
	Avg. rank	20.76	19.11	

Note. p-value exact. TA = Task achievement, FLIN = Fluency, RSL = Range of spoken language, ASL = Accuracy of spoken language.

Accuracy. Table 5.16 presents the intercorrelation matrix from correlating measures of accuracy with Rationality and Experientiality scores. As can be seen

from the table, there were no significant relationships between processing mode and rating accuracy. All coefficients were below +/- .2 and indicate that there is likely no relationship between the variables. The highest positive coefficients were between Rationality and accuracy on the criteria FLIN and RSL (both $r_s = .17$, $p = .30$ and $p = .30$). The lowest negative coefficient emerged between Rationality and accuracy on TA ($r_s = -.13$, $p = .44$).

Table 5.16 Spearman correlation between measures of accuracy and preferred cognitive processing mode (N = 39)

Variable	Rationality	Experientiality
Accuracy (full model)	.05 [-.28, .39]	.04 [-.29, .35]
Accuracy (TA)	-.13 [-.41, .16]	-.08 [-.38, .25]
Accuracy (FLIN)	.17 [-.16, .49]	-.10 [-.45, .25]
Accuracy (RSL)	.17 [-.16, .49]	.09 [-.26, .41]
Accuracy (ASL)	.01 [-.32, .35]	.15 [-.19, .45]

Note. TA = Task achievement, FLIN = Fluency, RSL = Range of spoken language, ASL = Accuracy of spoken language. Values in square brackets indicate the 95% confidence interval for each correlation.

A Mann-Whitney U test was run to identify if there were differences in rating accuracy between predominantly rational and experiential processors. Distributions of the accuracy scores in all five models appeared similar for rational and experiential raters when visually examining clustered histograms. Even though there were some noticeable differences in medians and mean ranks for the criteria ASL, FLIN and TA, there were no statistically significant differences between rational and experiential raters in any of the accuracy models (see Table 5.17). The null hypothesis that both groups rated similarly in terms of accuracy was retained.

Table 5.17 Summary of Mann-Whitney U tests between rational and experiential processors and rating accuracy

		Rational processors (N = 21)	Experiential processors (N = 18)	Exact <i>p</i>- value
Accuracy (full model)	Mdn.	1.57	1.53	.945
	Avg. rank	20.14	19.83	
Accuracy TA	Mdn.	1.49	1.45	.394
	Avg. rank	21.48	18.28	
Accuracy FLIN	Mdn.	1.96	1.76	.443
	Avg. rank	21.33	18.44	
Accuracy RSL	Mdn.	1.57	1.52	.967
	Avg. rank	19.9	20.11	
Accuracy ASL	Mdn.	1.46	1.74	.294
	Avg. rank	18.21	22.08	

Note. TA = Task achievement, FLIN = Fluency, RSL = Range of spoken language, ASL = Accuracy of spoken language.

5.4.3 Preferred Cognitive Processing Mode and Rating Behaviour

Finally, the Experientiality and Rationality score were correlated with the three measures of rating behaviour. As can be seen from the summary matrix (Table 5.18), there were no statistically significant correlations. The highest positive correlation was found between Rationality and Time to first decision ($r_s = .12$). The lowest correlation coefficient was found between Experientiality and Deliberation time ($r_s = -.20$), thus there was a weak association between experiential processing and being faster in taking rating decisions.

Table 5.18 Spearman correlations between measures of rating behaviour and preferred cognitive mode ($N = 39$)

Rater behaviour	Rationality	Experientiality
Deliberation time	.04 [-.32, .36]	-.20 [-.47, .09]
Time to first decision	.12 [-.22, .42]	.03 [-.28, .33]
Revisions	-.07 [-.40, .28]	-.08 [-.40, .24]

Note. Values in square brackets indicate the 95% confidence interval for each correlation.

5.5 General Decision-making Style Inventory (GDMSI)

5.5.1 Descriptive Statistics and Scale Properties

The descriptive statistics for the General Decision-Making Style Inventory (GDMSI) are presented in Table 5.19. The decision-making styles (DMS) measured through the *rational*, *intuitive*, and *dependent* subscales were endorsed more overall than the two subscales for *avoidant* or *spontaneous* DMS. The range between the minimum and maximum mean score was particularly large for the *avoidant* scale and rather narrow for the *rational* scale. Two variables, *spontaneous* and *intuitive* DMS were not normally distributed.

Table 5.20 presents the intercorrelation coefficients between the various subscales and their Cronbach's Alpha. The reliability coefficients from $\alpha = .71$ to $\alpha = .89$ are acceptable for all five DMS subscales and confirm an adequate degree of internal consistency (Green, 2013). The correlation coefficients mostly indicated no or only very weak associations between the different constructs. Noteworthy are two weak negative associations between the *rational* DMS and the *intuitive* ($r_s = -.30, p = .06$)

and *spontaneous* ($r_s = -.34, p = .03$) DMS, as well as the weak positive association between *spontaneous* and *avoidant* style ($r_s = .36, p = .02$). By far the strongest positive correlation was found between the *avoidant* and the *dependent* style ($r_s = .53, p < .01$). The patterns in terms of orientation and strength of most inter-correlational relationships in this sample were similar to those reported by the original authors of the scale on a sample of undergraduate students (Scott & Bruce, 1995). The scale properties are in line with expectations as the constructs captured by the various subscales are not mutually exclusive and participants can simultaneously score high or low in several styles.

Table 5.19 Descriptive statistics of General Decision-Making Style Inventory (GDMSI)

Variables	Min	Max	M	SD	Skewness	Kurtosis
Rational ($k = 5$)	2.60	4.80	3.87	0.51	-0.29	0.07
Intuitive ($k = 5$)	2.00	5.00	3.77	0.56	-0.95	1.82
Dependent ($k = 5$)	2.20	5.00	3.77	0.73	-0.40	-0.42
Avoidant ($k = 5$)	1.00	4.80	2.71	0.86	0.49	0.56
Spontaneous ($k = 5$)	2.00	4.40	2.84	0.64	0.54	-0.53

Note. $N = 39$. *SE for Skewness* = 0.38; *SE for Kurtosis* = 0.74.

Table 5.20 Spearman intercorrelations and reliabilities for the General Decision-Making Styles Inventory (GDMSI)

Decision making style	1	2	3	4	5
1. Rational	(.71)				
2. Intuitive	-.30	(.83)			
3. Dependent	.10	.23	(.84)		
4. Avoidant	-.20	.19	.53**	(.89)	
5. Spontaneous	-.34	.24	.11	.36*	(.80)

Note. $N = 39$. Reliabilities for the GDMSI scales (Cronbach's Alpha) appear in parentheses along the diagonal. * Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed).

5.5.2 Decision Making Style and Rating Quality

The GDMSI produced five independent variables (i.e., one variable for each DMS, see also item map in Appendix K.2) which were correlated with each set of measures of rating quality (severity and accuracy), and rating behaviour metrics (deliberation time, time to first decision and revisions).

Severity. As can be seen from the correlation matrix in Table 5.21, the analysis revealed a statistically significant agreement between the ranking of four measures of severity and two decision making styles. These were between the *intuitive* DMS and severity RSL ($r_s = -.48, p < .01$), as well as between the *avoidant* DMS and severity overall (full model, $r_s = -.52, p < .01$), severity RSL ($r_s = -.51, p < .01$), and severity ASL, ($r_s = -.50, p < .01$). All coefficients indicate a moderate negative association, and the null hypotheses can be rejected. The findings suggest that raters who preferred the *intuitive* or *avoidant* DMS were less severe when rating the criterion RSL, which is concerned with the range of structures a speaker employs in their performance. Preferring the *avoidant* DMS was also associated with less severity when rating speaker accuracy. As overall severity was related more strongly to the ratings on these two criteria (see inter-correlations in Table 4.9), the effect appears to have carried over to the association between the *avoidant* DMS and the overall severity measure. Two other noteworthy features are the orientation of coefficients, which appear to be positive only with the *rational* DMS, and the very weak associations between the *spontaneous* style and severity.

Table 5.21 Correlations between measures of severity (MFRM) and decision-making style (DMS)

DMS	Severity (full model)	Severity (TA)	Severity (FLIN)	Severity (RSL)	Severity (ASL)
Rational	.28 [-.01, .53]	.24 [-.05, .51]	.33 [.03, .58]	.16 [-.13, .42]	.21 [-.10, .49]
Intuitive	-.39 [-.64, -.10]	-.24 [-.52, .07]	-.32 [-.58, -.03]	-.48** [-.71, -.19]	-.40 [-.07, .56]
Dependent	-.38 [-.65, -.03]	-.36 [-.62, -.04]	-.27 [-.58, .09]	-.39 [-.64, -.08]	-.33 [-.61, .00]
Avoidant	-.52** [-.73, -.25]	-.38 [-.65, -.09]	-.41 [-.64, -.12]	-.51** [-.72, -.23]	-.50** [-.72, -.24]
Spontaneous	-.16 [-.45, .17]	-.03 [-.35, .29]	-.11 [-.38, .17]	-.15 [-.46, .19]	-.19 [-.49, .15]

Note. Note. TA = Task achievement, FLIN = Fluency, RSL = Range of spoken language, ASL = Accuracy of spoken language. Values in square brackets indicate the 95% confidence interval for each correlation. *p*-values corrected via Holm-Bonferroni stepwise procedure. ** Correlation is significant at the 0.01 level (2-tailed).

As the different decision-making styles are not mutually exclusive and raters may prefer both, the *intuitive* and the *avoidant* DMS at the same time, regression analyses were carried out to clarify the unique contributions of each decision-making style to severity when assessing the criteria RSL and ASL.

DMS and severity RSL. First, the relative contribution of the *intuitive*, *dependent*, and *avoidant* DMS, which all displayed strong correlations on rating severity for the criterion RSL, was investigated. A Durbin-Watson statistic of 2.093 indicated an independence of residuals. Linearity and approximate homoscedasticity were assessed by inspecting the scatterplot of studentized residuals against the predicted values. Linearity between the independent variable and each single dependent variable was established by inspecting partial regression plots. With no correlation coefficient exceeding .60 and all tolerance values above .691 or higher, there was no indication of collinearity in the data set. The data set was also investigated for outliers, leverage points and influential points. None of the studentized deleted

residuals was an outlier and greater than ± 3 standard deviations. Only three out of 39 leverage values were slightly above the recommended threshold of .2, with the highest value at .26. An inspection of Cook's Distance values showed no influential points in the data set that were above 1. A Q-Q Plot was used to assess the normal distribution of standardized residuals.

Next, following a backwards stepwise entry method all independent variables (*intuitive*, *dependent*, and *avoidant* DMS) were entered into the model. After the first step, the *dependent* DMS was dropped from the first full model as it did not significantly contribute to the regression model. R^2 for the full model 1 was 39.2% with an adjusted R^2 of 34.0% and 38.2% with an adjusted R^2 of 34.8% for the reduced model 2, a small to medium effect size (Cohen, 1988). In the regression model 1, two of the three variables contributed statistically significantly to the dependent variable (severity RSL), $F(3, 35) = 7.525, p < .005$. In model 2, the F -test improves slightly, $F(2, 36) = 11.136, p < .00$ (see summary Table 5.22).

The impact of the two decision making styles on severity (RSL) is considerable. In model 2, an increase in *intuitive* DMS as measured in the GDMSI by one unit is associated with a decrease of .39 logits in severity (RSL). Similarly, an increase of one unit in the *avoidant* DMS scale is associated with a decrease of .23 logits in severity (RSL). The standardized beta-values suggest that if the other variable be held constant, the *intuitive* DMS has a slightly stronger effect on severity in rating RSL ($\beta = -.42$) than the *avoidant* DMS ($\beta = -.38$).

Table 5.22 Summary of multiple regression analysis for severity (RSL)

	Variable	B	SE B	β
Step 1	Constant	2.25	0.53	
	Intuitive DMS	-0.37	0.13	-.39**
	Dependent DMS	-0.09	0.12	-.12
	Avoidant DMS	-0.20	0.10	-.32*
Step 2	Constant	2.10	0.48	
	Intuitive DMS	-0.39	0.13	-.42**
	Avoidant DMS	-0.23	0.08	-.38**

Note. * $p < .05$, ** $p < .01$; B = unstandardized regression coefficient; SE B = Standard error of the coefficient; β = standardized coefficient

DMS and severity ASL. The contribution of the *intuitive*, *dependent*, and *avoidant* DMS were also investigated for severity in the criterion ASL. Independence of residuals was indicated by a Durbin-Watson statistic of 2.544. Scatterplots of studentized residuals against unstandardized predicted values were inspected for approximate homoscedasticity and linearity between the dependent variable and all predictor variables. Partial regression plots indicated linear relationships between each independent variable (DMS) and the dependent variable (severity ASL). Collinearity could be excluded based on the tolerance values and low to moderate correlation coefficients. There were no outliers with no studentized deleted residual exceeding a standard deviation of ± 3 , only three leverage points slightly above .2, and no Cook's Distance value above 1. A normal distribution of standardized residuals could be confirmed through a Q-Q Plot.

From a backwards stepwise regression, the *intuitive* and *avoidant* DMS emerged to be contributing statistically significantly to the outcome variable (severity ASL), model 1 $F(3, 35) = 6.033, p < .005$; reduced model 2 $F(2, 36) = 9.15, p < .005$. R^2 for model 1 was 34.1%, with an adjusted R^2 of 28.4%; for model 2 R^2 was 33.7%

with an adjusted R^2 of 30.0%. As can be seen in the summary (Table 5.23), in the improved model 2 an increase in *intuitive* DMS by one unit was associated with a decrease of .36 logits and a similar increase by one unit for the *avoidant* DMS leads to a decrease of .32 logits in severity ASL. According to the standardized beta values, the *avoidant* DMS has a slightly stronger effect on severity (ASL) ($\beta = -.32$) than the *intuitive* DMS ($\beta = -.42$).

Table 5.23 Summary of multiple regression analysis for severity (ASL)

	Variable	<i>B</i>	<i>SE B</i>	β
Step 1	Constant	2.35	0.67	
	Intuitive DMS	-0.35	0.17	-.30*
	Dependent DMS	-0.07	0.15	-.08
	Avoidant DMS	-0.29	0.12	-.39*
Step 2	Constant	2.24	0.61	
	Intuitive DMS	-0.36	0.16	-.32*
	Avoidant DMS	-0.32	0.11	-.42**

Note. * $p < .05$, ** $p < .01$; *B* = unstandardized regression coefficient; *SE B* = Standard error of the coefficient; β = standardized coefficient

Accuracy. The correlation matrix for the associations between decision making style and all MFRM-based rater accuracy variables can be found in Table 5.24. Overall, correlation coefficients range from indicating no association at all ($r_s = -.00$; for *spontaneous* DMS with the full model) to moderate ($r_s = .53$, $p = .01$; for *avoidant* DMS with accuracy in RSL). Only the association between the *avoidant* DMS and accuracy in RSL was statistically significant after the post hoc Holm-Bonferroni correction of p -values. The coefficient reflects a significant overlap between the ranking of values on the two criteria and indicate that raters who responded as more avoidant were also more accurate in their rating (in addition to being somewhat less severe, as described above).

Apart from this finding, there are a few noteworthy patterns in the data. The *rational* DMS generally behaved differently to the other four DMS and produced negatively oriented coefficients. Furthermore, three decision making styles (*intuitive*, *dependent*, and *avoidant*) tended towards higher coefficients for the language related criteria RSL and ASL. This was particularly the case for accuracy on the criterion RSL. In comparison, coefficients for accuracy in the criteria FLIN and TA were generally smaller and appeared less associated with responses on the GDMSI.

Table 5.24 Correlations between measures of accuracy and decision-making style (DMS)

DMS	Accuracy (full model)	Accuracy (TA)	Accuracy (FLIN)	Accuracy (RSL)	Accuracy (ASL)
Rational	-.14 [-.47, .21]	-.09 [-.41, .25]	-.09 [-.42, .27]	-.14 [-.46, .20]	-.14 [-.44, .18]
Intuitive	.29 [-.02, .56]	.02 [-.29, .35]	.16 [-.13, .44]	.42 [.10, .68]	.40 [-.07, .66]
Dependent	.31 [-.03, .59]	.13 [-.24, .47]	.15 [-.14, .44]	.40 [.11, .64]	.29 [-.02, .56]
Avoidant	.37 [.09, .60]	.09 [-.25, .42]	.25 [-.07, .54]	.53** [.28, .72]	.31 [.03, .54]
Spontaneous	.00 [-.35, .34]	-.14 [-.46, .19]	.04 [-.30, .36]	.10 [-.26, .44]	.03 [-.33, .37]

*Note. Note. TA = Task achievement, FLIN = Fluency, RSL = Range of spoken language, ASL = Accuracy of spoken language. Values in square brackets indicate the 95% confidence interval for each correlation. p-values corrected via Holm-Bonferroni stepwise procedure. ** Correlation is significant at the 0.01 level (2-tailed).*

As was the case for severity, *intuitive*, *dependent*, and *avoidant* DMS showed small to moderate correlations with rating decisions concerning the language related criteria RSL and ASL. As there was a significant correlation of DMS with accuracy for RSL, a regression analysis was conducted to investigate the possible contribution of these variables to accuracy RSL.

DMS and accuracy RSL. The Durbin-Watson statistic (2.243) confirmed that residuals were independent. Linearity between each predictor variable (*intuitive*, *dependent*, and *avoidant* DMS) and the dependent variable (accuracy RSL) was confirmed by inspecting scatterplots. Similarly, overall linearity between all dependent variables and the dependent variable as well as homoscedasticity were investigated through a scatterplot of studentized residuals against unstandardized predicted values. There was no evidence of collinearity in the tolerance values and correlation coefficients. All studentized deleted residuals remained within +/- 2 standard deviations, there were only three leverage points slightly above .2, and no Cook's Distance value above 1. A Q-Q Plot of standardized residuals displayed normal distribution.

A backwards stepwise regression was conducted which showed that the *intuitive* and *avoidant* DMS are contributing statistically significantly to the outcome variable (accuracy RSL). Model 1 had a R^2 value of 35.7% with an adjusted R^2 of 30.2% while the R^2 in model 2 was 35.0% and the adjusted score 31.4%. Both, *intuitive* and *avoidant* DMS contributed statistically significantly to accuracy when rating RSL, model 1 $F(3, 35) = 6.477, p < .005$; reduced model 2 $F(2, 36) = 9.705, p < .001$ (see Table 5.25). The positive β -values indicate a positive relationship of the predictor variable with the outcome variable. Thus, in the improved model 2 an increase in *intuitive* DMS by one unit is associated with an increase of .31 logits in accuracy and a similar increase by one unit for the *avoidant* DMS can be expected to lead to an increase of .24 logits in accuracy for RSL. The standardised beta-values further suggest that the *avoidant* DMS emerges to have a slightly higher impact on accuracy RSL ($\beta = .41$) than the *intuitive* DMS ($\beta = .35$).

Table 5.25 Summary of multiple regression analysis for accuracy (RSL)

	Variable	<i>B</i>	<i>SE B</i>	β
Step 1	Constant	-0.36	0.52	
	Intuitive DMS	0.29	0.13	.33*
	Dependent DMS	0.07	0.11	.10
	Avoidant DMS	0.21	0.09	.36*
Step 2	Constant	-0.24	0.47	
	Intuitive DMS	0.31	0.12	.35*
	Avoidant DMS	0.24	0.08	.41**

Note. * $p < .05$, ** $p < .01$; *B* = unstandardized regression coefficient; *SE B* = Standard error of the coefficient; β = standardized coefficient

5.5.3 DMS and Rating Behaviour

Variables of decision-making style were correlated with the three rating behaviour metrics (deliberation time, time to first decision, revisions). The analysis failed to produce statistically significant results. When taking the confidence intervals into account, the two variables measuring duration (i.e., deliberation time and time to first decision) produced patterns that could be expected for the *rational* and the *spontaneous* DMS as there was a tendency for raters preferring the *spontaneous* DMS to take less time for the rating overall and for entering their first decisions. Raters preferring the *rational* DMS, on the other hand tended to take longer overall and entered their first decision later. Furthermore, there was a positive correlation between the *avoidant* and *dependent* DMS and number of revisions, which coincided with expectations as raters with a preference for these styles tended to revise their decisions frequently compared to raters preferring other DMS.

Table 5.26 Spearman correlations between rater behaviour metrics and decision-making styles (DMS)

Rater behaviour	Rational	Intuitive	Dependent	Avoidant	Spontaneous
Deliberation time	.37 [.10, .60]	-.07 [-.39, .28]	.15 [-.15, .41]	.03 [-.25, .31]	-.36 [-.60, -.08]
Time to first decision	.27 [-.03, .52]	-.02 [-.32, .35]	-.15 [-.41, .11]	-.13 [-.42, .18]	-.23 [-.54, .10]
Revisions	.05 [-.25, .36]	-.06 [-.39, .28]	.29 [-.04, .57]	.25 [-.07, .55]	.05 [-.54, .10]

Note. TA = Task achievement, FLIN = Fluency, RSL = Range of spoken language, ASL = Accuracy of spoken language. Values in square brackets indicate the 95% confidence interval for each correlation. *p*-values corrected via Holm-Bonferroni stepwise procedure.

5.6 Summary

In this chapter, the results for the exploratory correlational analyses between the dependent variables of rater quality and rater behaviour metrics, and the independent variables were presented (for summary tables see Appendix O). This study aimed at answering the second research question: *When novice raters assess speaking, are rating quality and rating behaviour metrics related to cognitive attributes, preferred processing mode or decision-making styles?*

The descriptive statistics of the three, time stamp-based rater behaviour metrics (deliberation time, time to first decision and revisions) were presented and their relationship to the dependent variables of rater severity and accuracy investigated. The number of revisions was moderately related to accuracy on the criterion TA. Other associations were found to be non-significant.

The correlational analyses investigating the relationships between the dependent variables and five cognitive measures (Stroop, Letters-Numbers, Digit Span, Keep Track, and Trail Making Task) produced no statistically significant results.

Similarly, no statistically significant results emerged from correlating the dependent variables with the two preferred cognitive modes (i.e., rational and experiential).

The correlational analyses of the associations between the rating quality and decision-making style (based on the GDMSI), however, revealed statistically significant correlations. The *avoidant* DMS correlated significantly and negatively with rater severity on the language related criteria RSL and ASL, as well as positively with rater accuracy on the criterion RSL. The *intuitive* DMS behaved similarly and correlated negatively with severity on RSL.

A series of regression analyses showed that the *avoidant* and *intuitive* DMS exerted significant, moderate effects on accuracy (ASL) and severity (ASL), with *avoidant* emerging as a slightly stronger predictor than the *intuitive* DMS. Severity (RSL) was also predicted by both DMS styles, but here the *intuitive* DMS had a slightly stronger effect.

Correlating DMS with rater behaviour metrics produced no statistically significant results after the Holm-Bonferroni correction. Correlational patterns, however, behaved in line with expectations as the *rational* DMS tended towards longer deliberation times and later first decisions, and the *spontaneous* DMS tended to work faster. *Avoidant* and *dependent* DMS on the other hand tended towards more frequent revisions.

6 Results III: Rater Case Studies

6.1 Outline

This chapter presents findings that address the third research question: *To what extent do case studies of accurate and inaccurate raters reveal differences in rating behaviour and influences on rater behaviour?* This focus on the individual participant was achieved by compiling four case studies based on individual measures of data previously presented at the group level in Chapters 4 and 5 and combining these with qualitative data gleaned throughout the rating process. Section 6.2 will introduce the four cases selected as case studies. Sections 6.3 to 6.6 present each case in detail. Section 6.7 will close this chapter with a between-case analysis which compares and contrasts the four case studies.

6.2 Selected Cases

The main selection criterion was rater accuracy as established by the MFRM RAM analyses. Before arriving at the final set, other attributes were also taken into consideration (for details see Methodology, Section 3.8.7). Stormi and Lexi were the first two cases identified and selected. They emerged as extremely capable raters as they were the most accurate overall. However, they were also slightly different in terms of other attributes such as decision-making style and cognitive preference. Deciding on two, lower accuracy raters was more challenging. Mary and Betty were eventually selected after taking into account their rater accuracy, the quality of their handwritten notes, their rating behaviour metrics and rater fit. They were both particularly inaccurate in their rating decisions despite a clear effort and

engagement evident from their handwritten notes and rating behaviour metrics. They differed, however, in the consistency of their ratings as Mary was more erratic and Betty was more consistently inaccurate.

Table 6.1 Key selection attributes for case studies

Low accuracy	High accuracy
<p>Mary</p> <ul style="list-style-type: none"> • erratic • rational and dependent DMS • rational processing 	<p>Stormi</p> <ul style="list-style-type: none"> • balanced rational and intuitive DMS • slight preference for experiential processing
<p>Betty</p> <ul style="list-style-type: none"> • consistent, restricted • tending towards <i>rational</i> DMS • slight preference for experiential processing 	<p>Lexi</p> <ul style="list-style-type: none"> • intuitive and spontaneous DMS • experiential processing

Before finalising the selection of cases, I investigated whether experience in terms of semesters of study or age were likely to have affected the main selection criterion, i.e., rater accuracy. A correlation analysis of rater accuracy on the full model as well as single criterion-based models with the independent variables *age* and *semester* revealed that these two variables are not associated with any of the dependent variables. For the variable *semester*, coefficients ranged from $r_s = -.24$ ($p = .14$ with accuracy TA) to $r_s = -.01$ ($p = .93$ with accuracy FLIN). For the variable *age*, correlation indices were lowest with accuracy RSL ($r_s = .01$, $p = .94$) and, in relative terms, highest with accuracy ASL ($r_s = -.08$, $p = .64$). No further investigation into the relationships between experience, as measured by the variables *age* and *semester*, and accuracy were carried out.

6.3 Case Study 1: Stormi

At the time of the data collection, Stormi was 23 years old which makes her considerably younger than the average rater in this study ($M_{age} = 25.2$ years). She was also less advanced in her studies as she was in her sixth semester ($M_{semester} = 8.6$).

Accuracy. Stormi emerged as the most accurate rater in terms of classical percentage agreement with the reference scores (60.00 %) and she was ranked among the top raters based on Cohen's weighted kappa ($k = .55$). Stormi scored lower in traditional measures of consistency, $r = .68$ and $\tau_b = .59$, than some of the other very accurate raters.

In the MFRM RAM analysis, Stormi emerged as the most accurate rater (logit = 2.15). She was also the only participant to perform exceptionally well across all four criteria, i.e., within one standard deviation of the maximum logit score attained by the cohort. She ranked second in the RAM model for ASL (2.39 logit, $SE = .27$), and third in the other three criteria (TA 2.00 logit, $SE = .24$; FLIN 2.58 logit, $SE = .28$; RSL 2.35 logit, $SE = .28$). Stormi did so without overfitting in any of the criterion-specific models, with MS_w ranging from .89 (TA) to 1.06 (ASL). This suggests that being highly accurate, Stormi was taking her decisions independently and without being too restricted.

Severity. Stormi was among the least severe raters overall (ranked 30 of 39, $-.32$ logit, $SE = .09$). In the criterion-specific severity models, her ratings rank 23rd, and in the middle of the group, for FLIN ($-.05$ logit, $SE = .20$), but within the top quartile

of the most lenient raters for the other three criteria (TA $-.48$ logit, $SE = .18$, rank 34, RSL $-.35$ logit, $SE = .19$, rank 30, ASL $-.64$ logit, $SE = .19$, rank 32).

Stormi's ratings were less erratic and, similar to the scores provided by four other raters, tended towards overfit ($MS_w = 0.68$) in the general RSM model that included all four criteria. In the criterion-specific analysis based on a reduced set of data, Stormi's ratings were overfitting for the criterion RSL ($MS_w = 0.68$), but at or above an MS_w of 0.7 for the other three criteria (TA $MS_w = 0.70$, FLIN $MS_w = 0.72$, ASL $MS_w = 0.77$).

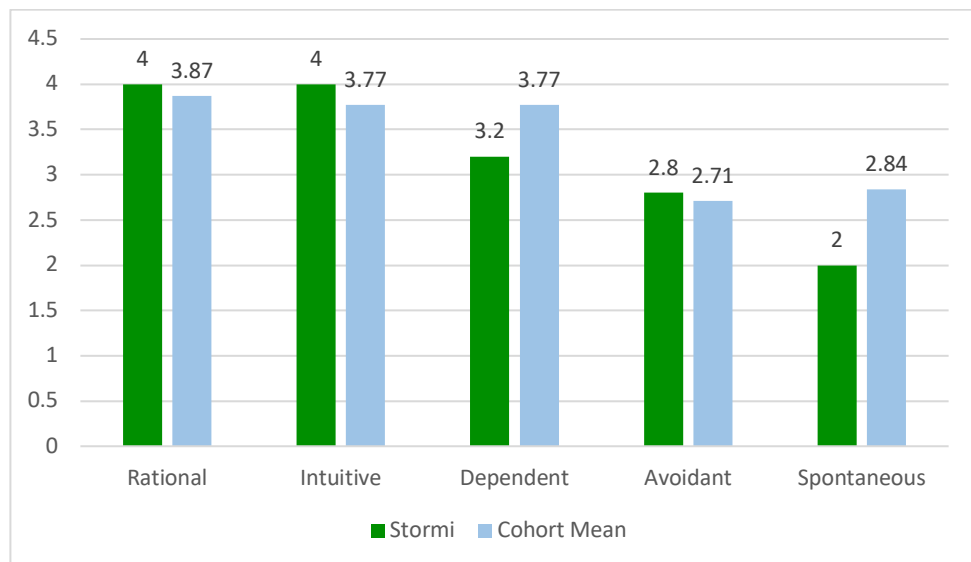
Thus, as far as rating quality is concerned, Stormi was a model rater compared to many of her peers in this study.

Cognitive variables. Stormi performed well on two of the five cognitive tasks. She ranked second for the Keep Track task and seventh for the Letters-Numbers task. Her scores on the Digit Span, Stroop test and Trail Making Task are in the middle of the field (rank 20, rank 24 and rank 17, respectively).

Cognitive preference. Stormi's scores on the REI-40 indicate that Stormi does not prefer either cognitive mode: rational nor experiential. Her mean experientiality score (3.7) was at the same level as the cohort. However, with a score of 4.0 she had an above-average preference for experiential engagement compared to the group ($M_{EE} = 3.7$). Stormi's overall rationality score (3.4) was just slightly below the average ($M_R = 3.66$). While she judged her rational ability at about the same level as the group level (3.5), her score of 3.3 indicates below-average engagement in rational processing ($M_{RP} = 3.77$). Her PSI score of 3 also indicates her to be a low-experiential processor.

Decision-making style. Stormi's profile with regards to the GDMSI showed a balanced orientation toward *rational* and *intuitive* DMS, which she preferred more than the cohort on average (see Figure 6.1). Her preference for the *dependent* DMS was slightly below that of the group and her score for the *avoidant* DMS was at about the group's average level. Perhaps the most noteworthy aspect is that Stormi clearly disfavours the *spontaneous* DMS and her score was almost a full point below the group mean ($M = 2.84$).

Figure 6.1 Stormi's DMS profile



Rating behaviour metrics. Despite a relatively low inclination towards spontaneity, Stormi was quite fast in taking her decisions. In both sessions as well as overall, she was among the top third in terms of speed to taking a decision. Stormi's average deliberation times ($M_{DT} = 29.28$) are clearly below the cohort average ($M_{DT} = 59.86$). Stormi was also fast when taking her first decision. On average, she would enter her first rating decision after 168.75 seconds in Session 1 and after 201.18 seconds in Session 2. This is considerably faster than the group means ($M_{TTFD \text{ Session1}} = 487.69$ and $M_{TTFD \text{ Session 2}} = 418.53$).

However, importantly, Stormi also tended to revise more than the average raters in the cohort. She is within the top third of revisers for Session 1 (ranked 11th) and top half for Session 2 (ranked 15th). In total, she made 29 changes, with most for ASL ($n = 9$), followed by TA ($n = 8$), RSL ($n = 7$) and FLIN ($n = 5$). These figures appear to be in line with Stormi's characterisation as a balanced decision-maker. On the one hand, she took initial decisions quickly and intuitively, but also opted to revise her decisions – in line with a more rational approach – based on new information as the performance continued.

Perception of rating. Stormi's self-report data concerning her confidence and perceived difficulty of rating remained stable across the rating sessions and four measurement points (middle and end of each session). After both sessions, Stormi indicated that she was "somewhat confident" (3 out of 5) about her rating decisions and felt that rating was "slightly difficult" (4 out of 5). As far as her confidence was concerned, her self-perception is similar to that of many other participants. The middle option, "somewhat confident", was selected by 46.2 to 48.7 % after the first three intervals. Only after having rated all 30 performances, did more participants (41.0 %) report to be "moderately confident" (4 out of 5 on the scale) than "somewhat confident" (35.9 %). Stormi also shares her perception of the difficulty of the rating process with most of her peers in the beginning of the experiment. However, while the majority shifts from "slightly difficult" (56.4 % and 53.8 % after the first 8 and 15 performances) towards a shared lead with the middle category "neither easy nor difficult" (43.6 % and 35.9 % each after 23 and 30 performances), Stormi's perception remains unchanged.

Stormi's comments regarding the difficulty of rating each criterion are presented in Table 6.2. Most of her responses were similar to that of the majority of the cohort. The greatest shift can be observed concerning the criterion TA. Like many other participants, Stormi found TA particularly difficult after the first rating session, where she rated this criterion to be "very hard", but she grew much more confident with the criterion in Session 2, where she rated it as "easy". After Session 1, Stormi problematized the difficulty of having to split her attention in order to be able to assess whether what participants said was "enough". However, her comment after Session 2 regarding TA seems to indicate that she settled on a more pragmatic approach to be lenient and focus on whether bullet points were addressed or not.

Table 6.2 Stormi's comments about the criteria

Criterion	Session 1	Session 2
	very difficult	easy
TA	It's not always clear what they are talking about; hard to keep track of what they are saying and decide if what they said was enough.	In my opinion, I was not very strict, if the points were addressed sufficiently it was enough for me
	difficult	difficult
FLIN	Hard to tell what a spontaneous element was, speed of talking is not the same as fluency for me, slow speakers can speak fluently too	Since there is no interaction, many descriptors cannot be applied
	easy	difficult
RSL	Easier in terms of deciding which band to pick, but difficult because the differences between the bands are not so clear to me (what is sufficient, what is wide)	Difficult to know what kind of lexicon the students have and what can be expected
	easy	easy
ASL	Grammar and pronunciation mistakes are usually easier to notice	Usually mistakes in that category are more apparent than others

Judging it to be “difficult” after both rating sessions, Stormi found the criterion FLIN more challenging to rate than was reported by the majority of the other participants. Stormi’s comment after Session 1 suggests that she struggled or did not fully agree with the concepts defined by the descriptors (“speed of talking is not the same as fluency for me”). After Session 2 she added that there were fewer useful descriptors in this criterion.

Like many of her peers, Stormi found rating RSL “easy” after Session 1 and “difficult” after Session 2. The main issue that Stormi appears to have with rating RSL is that she felt she had not yet developed an understanding of what constituted the desired proficiency level in terms of breadth (“what can be expected”). It is unclear from her comments why she felt the rating of RSL to be more difficult in Session 2.

Finally, Stormi found the criterion ASL quite easy to rate in both sessions. She justifies this by pointing out that mistakes in this criterion “are usually easier to notice” (S1) or “more apparent” (S2).

There are several underlying themes visible in Stormi’s comments regarding the different criteria. First, Stormi appears to engage quite intensely with the scale in Session 1. Her comments indicate that she is grappling with interpreting and applying the scale descriptors (problematizing the notion of a “spontaneous element”, and critiquing specific wording within the scale, “what is sufficient, what is wide”). In Session 2, Stormi appears more confident. For instance, she makes it clear that she opted for a pragmatic approach when dealing with the criterion TA, suggesting that she is willing to take independent decisions in order to deal with challenges posed by the scale descriptors. This shift is also visible in her comment

regarding RSL. In Session 1, Stormi seems to focus on the scale wording and her difficulty of identifying “differences between the bands” while her comment after Session 2 focuses on her grasp of the construct itself (i.e., “what can be expected”) rather than how it is operationalized in the scale.

Figure 6.2 Stormi's notes on P01 (rated 9th in Session 1 with 8 for TA and 7 for FLIN, RSL, and ASL)

Performance:			
TA	FLIN	RSL	ASL
Justification: th (something), stress (donating), treatment of food, consider the feeling ..., in poor living conditions, p vs. b (provided), inaccuracy concerning pronunciation, it's clear though what she's trying to say; how less food, provide awareness			

Figure 6.3 Stormi's notes on P04 (rated 6th in Session 2)

Performance:		P04	
TA	FLIN	RSL	ASL
9	7	7	6
Justification: - fillers (ähm, ...), cafeteria (pronunciation), clothes, avoid these current problem, biowaste? make an effort → frequent different errors concerning stress; awareness, chemist education - key difference, abundance, contribute strongly, nicely done			

Notetaking in rating forms. Figure 6.2 and Figure 6.3 are excerpts from Stormi's notes from rating performance 01 in Session 1 and performance 04 in Session 2⁷. Several aspects are striking about her notes. First of all, from the beginning onwards, Stormi summarizes her observations rather than just writing down direct quotes from the performances when she comments on the quality of *th*-sound or word stress. When looking through all of Stormi's notes, it also becomes clear that the breadth of features she observed was broad and included aspects such as pronunciation of individual sounds (*th*, *v-f*, *d-t*), stress, word order, complexity of sentence structure (if-clauses, relative clauses), circumlocution, overall fluency, whether errors were corrected, and quality of vocabulary. However, it is interesting that Stormi does not comment on TA in the first set of performances and systematic observations regarding this aspect only fade in towards the end of Session 1.

Another striking feature is that Stormi appears to have adapted her note taking style between Session 1 and 2. While her notes appear like a string of comments in Session 1, she structures her observations into strengths and weaknesses throughout the entirety of Session 2, which may indicate a shift in her own thinking towards a more precise representation of both her task of rating and the construct underlying the scale. During each rating session, there is no readily discernible variation in how

⁷ All notes included in this chapter are based on the two same performances, P01 from the second half of the first rating session and P04 from first half of the second rating session. These were chosen to make the notes at least somewhat comparable across the four raters. These performances were not associated with any bias on the group level MFRM analyses and approximately at the same ability level in terms of the fair scores based on the expert ratings. P01's fair scores were TA 8, FLIN 7, RSL 8, ASL 7, and P04's fair scores were TA 8, FLIN 8, RSL 8 and ASL 7.

Stormi takes and organises her notes. Rather, she seems to have adapted one strategy and maintained it from there on out.

Finally, a highly interesting feature of her notes is that Stormi documented some observations she has made about the students and how she interpreted them (e.g., “student laughs instead of finishing sentence, maybe lack of vocab?”). Given that Stormi did not take a lot of time to deliberate her ratings after the performances, she seems to have managed to interpret her observations and put them into words while the performances were still ongoing. At certain points, this feature also lends her notes a tone of confidence or authority; for instance, when she writes “inaccuracy concerning pronunciation, it’s clear though what she’s trying to say”.

Response to exit question. When responding to the open question in the interview, Stormi showed that she had a critical stance towards her performance as a rater. She described her approach as “naïve” and found it difficult to be fair. She would have liked to change some of her ratings later on:

“because you do not have a direct comparison and then you don’t know whether it makes any sense just based on this scale. It is difficult to stay fair because you might place someone lower on the scale in the beginning and then you hear someone else that seems to be at the same level but has other features that are better or worse. [...] you know that the student is at some level, at least that was my naïve understanding of the whole thing.”

She also does not fully subscribe to the features described in the scale concerning pronunciation. Instead, Stormi argues that what she considers minor issues with pronunciation would not interfere with intelligibility: “I would understand him, if he had difficulties with *b* or *p* sound and does not pronounce them as nicely. I’ll still know what he is trying to say.”

Summary. Stormi was selected as a case because she demonstrated an exceptional level of rating accuracy. She also displays a balanced profile with respect to cognitive tasks, cognitive preferences and decision-making styles. On the one hand, Stormi took a pragmatic approach and recorded her first decisions early on and quickly. However, she was willing to revise her decisions as the performances progressed and did so frequently.

Stormi's case appears to exhibit the very essence of a balanced approach, combining rational and structured decision-making with quick intuitive judgement. Stormi's notes reveal that she was able to develop and apply a rich representation of the rating scale to the performances as she noticed a broad range of features. She appears to have continued to develop her understanding of the scale further over the course of the rating sessions. Like other novice raters in this study, Stormi admitted that she struggled with certain descriptors and with paying attention simultaneously to content as well as language during the rating process. This appears to be symptomatic for novice raters who are new both to the scale and to the typical test performances they may encounter. Yet, Stormi balances out possible shortcomings of training or scale wording through actively searching for and defining her own solutions. Stormi's highly accurate set of ratings indicate a deep engagement with the scale and a consistent interpretation of it. While remaining flexible and vigilant, she keeps an eye on the overall impression of a performance and is not easily drawn towards local features like slips or isolated weaknesses. In this sense, Stormi's profile might best be characterised as *balanced-independent*.

6.4 Case Study 2: Lexi

At the time of data collection, Lexi was 27 and slightly older than the average rater ($M_{age} = 25.3$). She was in her eleventh semester of studies, which is also clearly above the average participant ($M_{semester} = 8.6$, $Mdn_{semester} = 7$).

Accuracy. Lexi was one of the most accurate raters. Her ratings ranked second in exact agreement (56.67 %) and obtained the highest weighted kappa ($k = .62$). Her ratings were even more consistent with the reference scores than Stormi's ($r = .80$, $\tau_b = .72$).

In the MFRM RAM analysis, Lexi's overall accuracy measure across all four criteria was estimated at 2.07 logits ($SE = 0.13$). With a total accuracy score of 638, she only scored five points below Stormi (total score = 643) and ranked second in the cohort. Lexi emerged as exceptionally accurate in the criteria FLIN (ranked 1st, 3.03 logits, $SE = 0.32$) and ASL (ranked 2nd, 2.39 logits, $SE = 0.27$), where she landed within the top standard deviation of the maximum score obtained in the cohort. Lexi just about missed being within the top standard deviation for RSL (2.05 logits, $SE = 0.26$). Lexi's ratings were the least accurate in the criterion TA. However, with an estimated accuracy measure of 1.68 logits ($SE = 0.22$), she still ranked above the mean group logit of 1.47. There were no significant rater-criterion interactions for Lexi's ratings, indicating that her accuracy was not significantly stronger or weaker for any of the criteria.

Severity. Overall, Lexi was more lenient than Stormi with -0.46 logit ($SE = 0.09$), and only five raters were more lenient than her. For the two criteria RSL (-0.57 logits, $SE = 0.19$) and ASL (-0.71 logits, $SE = 0.19$), Lexi was among the most

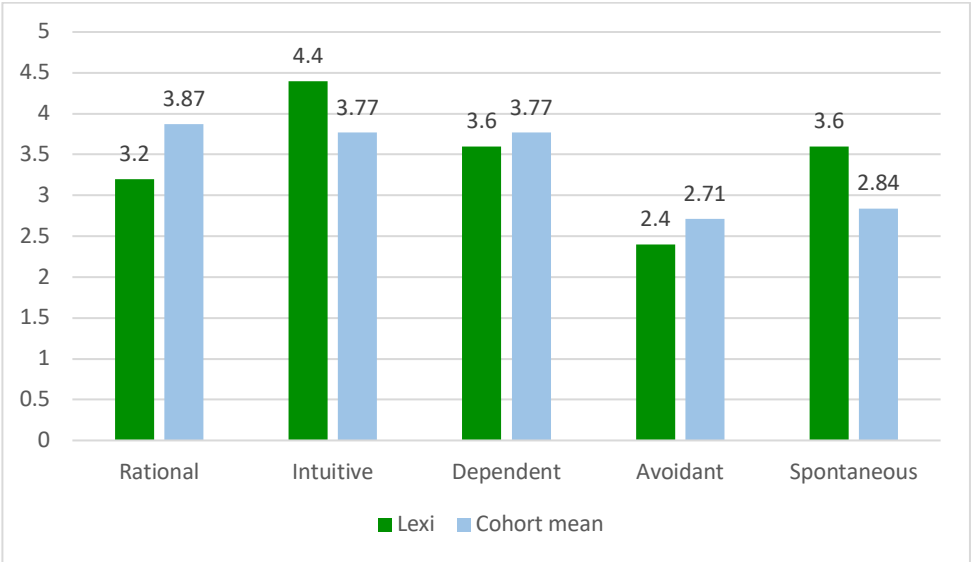
lenient raters within one standard deviation of the minimum logit scores. For the other two criteria, she ranked 8th (TA, -0.48 logits, $SE = 0.11$) and 10th (FLIN, -0.50 logits, $SE = 0.21$). As with Stormi, there was a slight tendency to overfit (full model, $MS_w = 0.7$, $t_w = -2.67$), particularly for the criterion FLIN ($MS_w = 0.66$, $t_w = -1.41$). Lexi produced one unexpected rating with a significant residual when she rated performance 27 two bands lower than expected (8 instead of 10). However, this performance may have been more difficult to rate as six other raters also provided TA ratings with significant residuals with performance 27.

Cognitive variables. Lexi's performance in the cognitive tasks varied considerably. She was excellent in the Stroop test (4th) and the Keep Track test (6th). Lexi also performed better than the average in the Forward Digit Span test (8, $Mdn = 7$). However, she was just slightly above average in Trail Making task (14th), and performed poorly on the Letters-Numbers Task (36th).

Cognitive preference. The data from the REI-40 questionnaire suggested that Lexi preferred experiential processing over rational processing. Her individual mean scores on the rational ability scale (3.9), rational engagement scale (4.1) and the overall rationality scale (4.0) were all clearly above the respective group means of 3.56, 3.77 and 3.66. Nonetheless, Lexi's scores were even higher for the experiential scales. She rated her experiential ability at 4.9 and her experiential preference at 4.7 which are both distinctly above the group means of 3.7. As a result, Lexi's experientiality score of 4.8 was higher than the group mean of 3.7. When combining the scores from all sub-scales into a PSI score ($PSI = 13$), Lexi emerges as a high-experiential processor. Her score is the most distinct preference for a processing style of all four case studies included in this analysis.

Decision-making style. While Stormi showed a balanced profile between the *rational* and *intuitive* DMS (see Section 6.3), Lexi clearly preferred the *intuitive* ($M = 4.4$) over the *rational* ($M = 3.2$) DMS (see Figure 6.4). This is in congruence with her PSI score. Another interesting feature is that she expressed the strongest preference for the *dependent* and *spontaneous* DMS (both $M = 3.6$) of all four cases. Lexi’s score for the *avoidant* style ($M = 2.4$) is clearly below the cohort mean ($M = 2.71$).

Figure 6.4 Lexi's DMS profile



Rating behaviour metrics. Lexi’s mean deliberation time ($M_{DT} = 24.05$ s) indicated that she was the swifter in completing her ratings than any of the four raters and the overall cohort mean ($M_{DT} = 60.16$ s, ranked 6th). However, she was not particularly fast compared to the whole group when taking down her first rating decision ($M_{TTFD} = 253.30$, ranked 15th) and also slower than Stormi ($M_{TTFD} = 184.96$ s) in this respect. As far as revisions are concerned, Lexi was in the middle of the field with a total of 18 revisions. Most of her revisions concerned RSL ($n = 8$), followed by TA ($n = 5$), FLIN ($n = 4$) and ASL ($n = 1$). Looking at the data collated

so far, Lexi appears to engage more intuitively with the rating task as she displays preferences for the *intuitive* and *spontaneous* DMS and spends little time overall to pore over the criteria and performances. However, the high accuracy of her scores and number of revisions indicate that she manages to engage successfully with the rating task, holds back on quick first decisions and is open to adjusting her impressions as the performance unfolds.

Perception of rating. Overall, Lexi reported that she was “moderately confident” (4 out of 5) after the first half of Session 1. However, she settled for “somewhat confident” (3 out of 5) after the first session and remained there for the rest of the experiment. This also happened to be the group mean for this variable ($M_{confidence} = 3.07$). Same as her peers, Lexi perceived rating to be somewhere between “neither easy nor difficult” (3 out of 5) and “slightly difficult” (4 out of 5, $M_{difficulty} = 3.5$).

As far as the difficulty of rating the individual criteria was concerned, Lexi’s responses were similar to that of the cohort for the criteria FLIN (2.5, $M_{difficulty\ FLIN} = 2.65$), RSL (2, $M_{difficulty\ RSL} = 2.32$), and ASL (2.5, $M_{difficulty\ ASL} = 2.37$) (see Table 6.3). The criterion RSL was “difficult” for Lexi in both rating sessions. After Session 1, she commented that she found it hard to choose a band when stronger and weaker features can be found in a performance. After the second session, Lexi foregrounded that it was difficult to discern or notice whether certain features (“good expressions”) that are mentioned in the rating scale were in fact in a performance. There is a slight difference between these justifications as the first focusses more on the scale and how it relates to a performance, while the second stresses the difficulty of identifying certain features as the performance unfolds.

Table 6.3 Lexi's comments about the criteria

Criterion	Session 1	Session 2
	Easy	Easy
TA	Ticking each bullet point, when done might have helped me to see whether the task was fulfilled or not. Even though some aspects might have lacked in some performances, they all, at least tried to cover all points.	With checking the bullet points it was quite easy to see whether the students fulfilled all tasks or not.
	Difficult	Easy
FLIN	Sometimes it is hard to estimate whether a pause is productive or not. Also spontaneity is hard to evaluate.	When there were no pauses that hindered communication, it was fine
	Difficult	Difficult
RSL	Sometimes the students used repetitions of one word, but did find varied formulations for other words and so I did not know how to exactly evaluate this.	Especially judging whether they used various expressions or good circumlocutions was hard
	Easy	Difficult
ASL	If communication was not hindered the students should have fulfilled this part.	Some points were mostly fine and other not, so it was hard to find a path in between

The criterion where Lexi's perception differed somewhat from the group's perception was TA. Lexi found rating this criterion easier ($M = 3$, "easy" after both sessions) than the group ($M = 2.37$). After both rating sessions, she seemed pragmatic in interpreting the criterion, but also somewhat structured in her approach when she refers to "ticking" or "checking" whether certain points were covered in the performances. She also acknowledged that the speakers tried to "cover all points" even though not all of them have addressed these aspects fully, which reflects her genuinely more lenient rating decisions.

While her perception of the difficulty of rating RSL remained the same across sessions, there were noticeable shifts in her perception of FLIN and ASL. Lexi rated the criterion FLIN as “difficult” in Session 1, where she commented that she found it hard to interpret the kinds of pauses and identify spontaneity, to “easy” in Session 2. Based on her experience, Lexi appears to have defined a somewhat simplified ‘rule’ (“when there were no pauses that hindered communication”) and built up the confidence to rely on noticing this discerning feature. Lexi’s understanding of the rating task seems to have shifted from actively seeking certain features and interpreting the scale descriptors towards applying her own heuristic based on her understanding of the rating task and scale.

When Lexi reported ASL to be “easy” after the first session, she justified this by there being a threshold feature, i.e., if certain mistakes would hinder communication, to identify those students above or below a band. ASL appeared more difficult to her in Session 2, where she seemed to struggle conciliating uneven features in the performance (“some were mostly fine and others not”) with the rating scale. So while some of Lexi’s comments point towards her creating certain decision rules, her comments also suggest that while pragmatic and intuitive she does not rely on these decision rules alone but keeps being engaged in taking balanced decisions that take different performance features into consideration.

Figure 6.5 Lexi's notes on P01 (rated 15th in Session 1)

bad quality (sound)

Performance: 15/01			
TA	FLIN	RSL	ASL
7	7	7	7
Justification:			
BP1 captures up → social status ✓ BP2 in BP3 ~ ✓			
main problem are , p.m.: provided, product, awareness			
notes adj./adv.? create/repairs slips			

Figure 6.6 Lexi's notes on P04 (rated 3rd in Session 2)

Performance: 8/04			
TA	FLIN	RSL	ASL
9	8	8	8
Justification:			
p.m. "lafetio", "autres" "assistance" BP1 ✓ BP2 ✓ BP3 ✓			
repair ✓ unnecessary info: variation...			
circumlocution ✓ contribute & reply pretty nicely done			

Notetaking in rating forms. Figure 6.5 and Figure 6.6 show Lexi's notes from Session 1 (P01) and Session 2 (P04). Similar to Stormi, Lexi does not write down many direct quotes from the performances. On average, Lexi's notes are brief, to the point and include one phrase or word per performance in Session 1 and slightly more in Session 2. She occasionally phrases or categorises her observations using her own words during the first session ("non-productive pauses → no further ideas"; "structured" when describing TA), but this rarely happens during the second rating session. In terms of the range of features appearing in her notes, Lexi included observations regarding fluency ("few pauses", "spontaneous", "fluent"), range

(lists of synonyms used or paraphrases), and accuracy (mainly whether there were repairs or issues with pronunciation). Her notes, thus, seem similar in breadth to Stormi's and Lexi includes aspects from each criterion. However, Lexi's notes are more difficult to interpret as there is little clarification whether and how an observation impacts the rating. Her comments also do not address each criterion in each performance. In fact, for more than half of the performances ($n = 18$), there are no comments clearly referring to FLIN or RSL. However, this lack of notes does not necessarily indicate fatigue. Even the later ratings provided in Session 2 span several bands for the same performance which suggests that Lexi still engaged deeply with the rating process and tried to differentiate between the different features of a performance.

Another similarity to Stormi is that, in the beginning, Lexi's notes do not consistently refer to the content or ideas presented in the performances. After about the first seven performances, Lexi consistently notes down at least one aspect concerning TA and from that point onwards it is the only criterion she never fails to include in her notes.

Over the course of the first rating session, Lexi develops a kind of shorthand and some of its features are visible in Figure 6.5 with ticks or waves after the abbreviation "BP" for bullet point. In Figure 6.6, there are even more comments such as "repair", "circumlocution" or "pron." combined with ticks or examples ("cafeteria").

Response to exit question. When responding to the open question after the cognitive tests, Lexi's response focused on the aspect of experience and that she felt she lacked routine and practice when it came to assessing student performances:

“For me, it was much easier when we rated the categories separately [during the training session]. I have only briefly seen these scales before in our practical courses and looking back I think that we did not spend enough time with them. Once I will start working at a school and have to do all this rating, I think I will be overwhelmed and it would be great to learn more about this during our studies so that we get more practice. I noticed myself that during the first session and the second half of the first session in particular, that I got very tired and stricter. I had to tell myself ‘okay, no, do not give band 6 but change it to band 7’ [...] I think the more often you train the better you can assess your own performance as a rater and that’s also good for the students in the long run.”

Lexi’s comment also suggests that she observed her own rating behaviour and felt the need to regulate her thinking, particularly when she became aware of fatigue. Knowing that Lexi was a very accurate rater, the metacognitive effort she invested in this process appears to have successfully counterbalanced her preference towards intuitive decision making.

SUMMARY. Similar to Stormi, Lexi is a highly accurate and consistent rater who displayed a thorough but at the same time pragmatic approach to rating. Lexi’s scores in the cognitive tasks varied considerably as she was in the top of the field in the Stroop and Keep Track tasks, but performed average or poorly in the other tasks. While Stormi was balanced, Lexi showed a distinct preference for experiential processing, and the *intuitive* and *spontaneous* DMS. This was triangulated by comparatively short deliberation times and quite brief notes. Overall, Lexi’s intuitive approach appears to be a strong factor in shaping her approach towards rating.

In addition to her intuitive thinking, however, there is evidence of flexibility and of metacognitive awareness. For instance, Lexi seems to have held back before taking

her first decision and remained open to making revisions whenever she felt the need for it which could be interpreted as self-regulation through metacognitive processes. Her comments regarding the criteria indicate that she had identified critical features in the scales that she used as signposts to support her decision making. However, Lexi's notes during rating show that she managed to build her judgment on a broad representation of the construct, and she was a highly accurate rater with a tendency to highlighting positive features in the performances. It is therefore likely that she did not rigidly adhere to her self-generated signposts but remained actively engaged with the broader construct throughout the rating sessions. In response to the exit question, Lexi admitted that she found rating challenging and tiring which might also explain her mixed performance in the cognitive tasks. She also stated clearly that she tried to regulate her impulses when she felt that her ratings were starting to drift. Thus, she seemed to be able to balance challenges and strategies to meet them in the course of the experiment.

6.5 Case Study 3: Mary

At the time of the data collection, Mary was 23 years old and just in her fifth semester of studies. She was therefore among the younger and least advanced students in the cohort ($M_{age} = 25.3$, $M_{semester} = 8.6$, $Mdn_{semester} = 7$).

Accuracy. Mary was one of the weakest raters in this cohort. Her ratings only reached 16.67 % exact agreement with the reference scores, ranked among the least consistent ($r = .62$, $\tau_b = .52$), and also produced the one of the lowest accuracy scores of 546 ($M_{accuracy\ score} = 593.6$). Mary's overall accuracy was estimated at 1.03 logits ($SE = 0.09$), which is half a logit below the group mean, 1.52 logits ($SE =$

0.11). Mary was in the lowest scoring group (i.e., within one standard deviation of the minimum accuracy score) for the criteria FLIN (1.19 logits, $SE = 0.21$), RSL (0.96 logits, $SE = 0.21$) and ASL (0.59 logits, $SE = 0.18$). For RSL, Mary was in fact the least accurate rater. For the criterion TA, Mary reached an accuracy estimate of 1.49 logits ($SE = 0.21$) which was just slightly above the group mean, 1.47 logits ($SE = 0.21$). Mary's increased accuracy on TA emerged as a significant rater-criterion interaction (+0.21 logits), as did her lower accuracy on ASL (-0.01 logits).

Severity. In the full model, Mary's estimated severity (0.40 logits, $SE = 0.08$) ranked 8th in the cohort. However, there is great variance between Mary's severity on individual criteria. Mary's estimate for TA (-0.41 logits, $SE = 0.18$) was close to Lexi's and Stormi's lenient logit scores (both -0.48, $SE = 0.18$). Her ratings were quite harsh, though, for the criterion RSL (ranked 10th, 0.42 logits, $SE = 0.19$) and among the harshest for ASL (ranked 1st, 1.25 logits, $SE = 0.19$) and FLIN (ranked 5th, 0.89 logits, $SE = 0.19$).

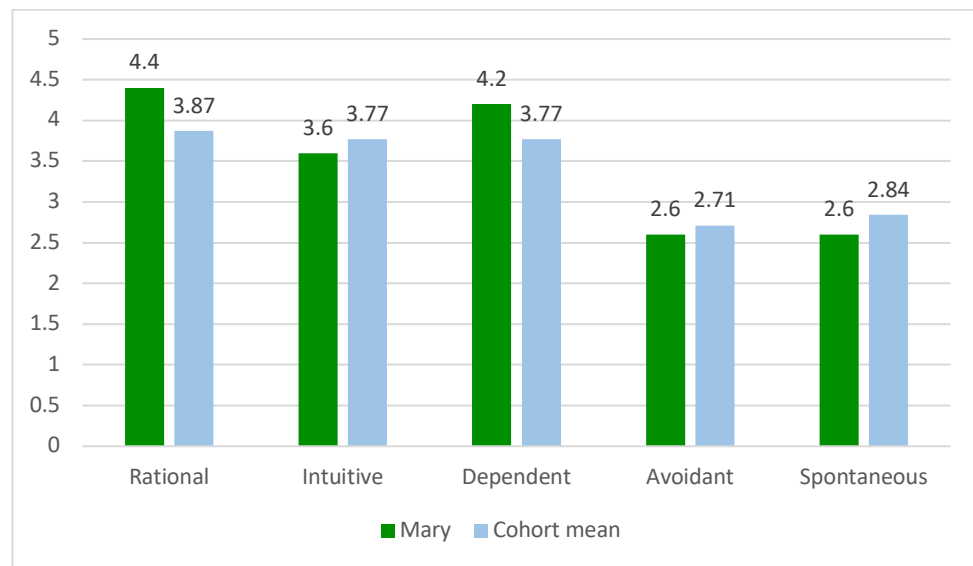
Mary's ratings were not only remarkable in terms of varying severity, but also associated with high mean-square measures in the overall model ($MS_w = 1.36$, $t_w = 2.56$), and the criteria RSL ($MS_w = 1.6$, $t_w = 2.07$) and ASL ($MS_w = 1.35$, $t_w = 1.33$). These indices suggest less predictable rating behaviour for the two language-focussed criteria, which appears to also have impacted the modelling of full model estimate. In the criterion FLIN, Mary's scores were more restricted than expected and tended towards overfit ($MS_w = 0.70$, $t_w = -1.14$). In the criterion TA, the mean-square index is within the optimum range of rater fit ($MS_w = 0.94$, $t_w = -0.15$).

Cognitive variables. Unlike Stormi and Lexi, Mary did not perform significantly above the average participant in any of the tasks. She obtained her highest score on the Keep Track test, where she ranked 16th (70.37, $M = 67.24$). All other scores were around or slightly below the group mean; Mary ranked 19th in the Stroop test (152, $M = 148.51$), 21st in the Letters-Numbers task (472, $M = 472.95$), 25th in the Trail Making task (66.27, $M = 67.77$), and was below the group median (7) in the Digit Span task (6).

Cognitive preference. Mary prefers rational processing over experiential processing. She scored 3.3 in the experiential ability subscale ($M_{EA} = 3.74$) and 3.5 in the experiential engagement subscale ($M_{EE} = 3.72$), leading to an overall experientiality score of 3.42 ($M_E = 3.73$). Her ratings for the rational processing scales were just slightly higher. Mary's average score on the rational ability was 3.3 ($M_{RA} = 3.56$) and on the rational engagement scale 3.9 ($M_{RE} = 3.77$). Overall, her rationality score (3.6) was about the same as the group mean ($M_R = 3.66$). Mary's PSI of -10 indicated her to be a high-rational processor.

Decision-making style. As can be seen in Figure 6.7, Mary's DMS profile shows a strong preference for the *rational* ($M = 4.4$) and *dependent* ($M = 4.2$) DMS as compared to the group. Her scores for the *intuitive* ($M = 3.6$), *avoidant* ($M = 2.6$) and *spontaneous* DMS ($M = 2.6$) are all below the cohort means.

Figure 6.7 Mary's DMS profile



Rating behaviour metrics. Mary took the longest of the four raters included as case studies for taking her decisions ($M_{DT} = 73.29$ s) and entering her first decision ($M_{TTFD} = 359.38$ s). In both variables she was considerably slower than the group ($M_{DT} = 60.16$, $M_{TTFD} = 262.30$). With a total number of seven revisions, she is also clearly below the group average ($Mdn_R = 18$). Most of her revisions concerned the criterion RSL ($n = 4$), followed by TA ($n = 2$), ASL ($n = 1$) and FLIN ($n = 0$).

The data collated thus far indicates that Mary appears to behave quite differently to the two previous cases. Her ratings show weaker rating quality and she generally obtained lower scores in the cognitive tasks, preferred a *rational* DMS and tended to take considerably longer and be less decisive (see *dependent* DMS) about her decisions.

Perception of rating. As can be seen from Table 6.4, Mary's perception of the difficulty of each criterion almost spanned the entire spectrum. Her judgement of the criteria TA and FLIN (both $M = 2.5$) was quite close to the cohort's average perception of these criteria ($M_{difficulty\ TA} = 2.37$, $M_{difficulty\ FLIN} = 2.65$).

Table 6.4 Mary's comments about the criteria

Criterion	Session 1	Session 2
	Easy	Difficult
TA	It can be observed well and checked by ticking every point.	I found it difficult to decide on performances where the interlocutor had to ask for additional points. Is the task still achieved, also if it wasn't done autonomously.
FLIN	Difficult	Easy
	It is often difficult, because it mostly changes during the performance.	I think it can mostly be observed right from the start of the performance how fluent a student is. However, it gets difficult if the fluency changes during the performance and they might don't know what to add anymore.
RSL	Difficult	Difficult
	Difficult because there are so many points to consider. But vocabulary in general is easy to assess.	It is often quite challenging to decide, since there are many aspects which have to be taken into consideration. For example some students have a wide range of vocabulary, but they do not use many complex structures.
ASL	Very difficult	Very difficult
	Very difficult, because there are many points to consider.	It was often very difficult for me, since I wasn't always able to pay simultaneously attention to pronunciation and grammatical structures. Moreover, I wasn't even always sure if some expressions even exist in English or not.

Concerning TA, Mary first perceived the criterion to be “easy” and applied a similar pragmatic approach as Lexi when she mentions “ticking every point” in her justification. However, after Session 2, Mary felt that this criterion was “difficult” and refers to certain performances where the interlocutor intervened to help participants fill the minimum speaking time. For the criterion FLIN, Mary first considered it to be “difficult” as this feature might fluctuate over the course of a performance. After the second round, Mary found rating this criterion “easy” and felt she could judge fluency “right from the start”, unless there were more severe

breakdowns in the speakers' delivery ("they might don't know what to add anymore"). She seemed to have found a way of dealing with the criterion and difficulties with rating this aspect were narrowed down to particular performances.

Mary generally felt that the two language-focussed criteria RSL and ASL were quite difficult to rate. In all of her statements concerning these two criteria, she stressed that the number of language features described in the scale posed a challenge for her. The criterion RSL was "difficult" (2) for Mary in both rounds, which is slightly below the group ($M_{difficulty\ RSL} = 2.32$). In addition to the number of descriptors, Mary also mentioned that choosing the right band in performances with mixed features posed a problem for her. She was one of two raters who found ASL "very difficult" at the end of the Session 2. While she argued this was due to the number of descriptors in the first round, she added in the second round that she struggled with paying "simultaneous attention" to different features in addition to not always knowing whether certain phrases were correct or not. ASL, thus, seemed to be challenging for this rater for several reasons. First, it is a complex criterion with a detailed description in the rating scale. In addition, Mary foregrounds the difficulty of having to switch or split her focus between certain language features (pronunciation and grammatical structures). Finally, Mary found it challenging to decide about the correctness of certain expressions in the real-time rating situation. This raises questions about Mary's English competence. Moreover, if she frequently engaged in trying to recall implicit or explicit knowledge about the correctness of certain expressions during the performance this might also have impacted on Mary's representation of the performances and increased cognitive load even more.

Notetaking in rating forms. Figure 6.8 and Figure 6.9 show Mary's notes from Session 1 (P01) and Session 2 (P04). What is striking about her notes in comparison to the two previously discussed raters, Stormi and Lexi, is the level of detail Mary managed to capture. Up to the last performance, each justification includes numerous observations about the performances and covers all four criteria.

Figure 6.8 Mary's notes on P01 (rated 12th in Session 1)

Performance: 12 (P01)			
TA	FLIN	RSL	ASL
6	5	8	4
Justification:			
<p>TA 1) description + differences ✓ 2) family ~ 3) school ✗</p> <p>FLIN "ein", pauses, uneven tempo</p> <p>RSL "wasn't needed", "charity organisation", "treatment of food"</p> <p>ASL the main problem are <u>pronunciation</u> "products", "purchase", "awareness" *</p> <p>④ satisfaction, social status, living condition, donate, food provided, grateful, nutrition</p> <p>* how poor people have to live ④ those nutrition → this nutrition</p>			
Rater training			

Figure 6.9 Mary's notes on P04 (rated 7th in Session 2)

Performance: 7 (P04)			
TA	FLIN	RSL	ASL
10	6	5	5
Justification:			
<p>TA contrast → "key difference" "family", schools ✓ → convincingly</p> <p>RSL 8.4 "is for throwing", to being eaten, giving away RSL 2.3</p> <p>ASL 5.4 <u>pronunciation</u> "clothes", "awareness" → restriction - searching for word</p> <p>FLIN 5.1 3-4 words - strings L switches to German! word "break" → pause</p> <p>RSL ④ apply to, ~ circumlocution "aesthetic?" ④ "chemist education"</p> <p>④ contribute strongly</p>			

For the criterion TA, which Mary commented on regularly and throughout both rating sessions, she included symbols such as ~ or ticks, but also took notes to

summarize the ideas or arguments (“contrast is missing”, “picture description is missing”).

For FLIN, there are references to the descriptors applied (e.g., “5.1” in Figure 6.9), quotes from the rating scale (“high degree of fluency and spontaneity”) and also descriptions in her own words (“slow but natural”, “long pauses → partly effective”). There is some difference between the words given in the rating scale and the words chosen by Mary to describe phenomena of fluency. While the scale does mention a lack of fluency as a feature of a less successful performance of band 4 or below, Mary awarded performances described as “slow but natural” or “clear and slow, natural” with band 6 and 9, respectively. In Session 2, Mary’s wording is closer to the wording of the descriptors in the scale. Her comments regarding the criterion RSL often include direct quotes from the performances (RSL “gets reused”, “freshly cooked”, “I don’t know the name”, “value food”, “give opportunity”, “expired”, “in need of”) or complex structures she noticed (if clauses, modals, passive).

Finally, for the criterion ASL, Mary also included many references to the performance often combined with some sort of correction or indication as to its accuracy (ASL sold → sold, “th” → throwing, the main problem are, “set a statement”).

Mary seems to develop the most notable features of her note-taking system early on and only adapts it minimally throughout the rating sessions. After the fourth performance in Session 1, Mary labels some of the lines with an abbreviation of each criterion and, thus, structures at least part of her notes. However, the clear order as seen in Figure 6.8 is not kept up for all of the performances and sometimes

it becomes difficult to tell which observation is dedicated or noted down for which criterion.

In comparison to the other two raters, Mary's notes can be said to be much richer in terms of examples from the performances. In comparison with Stormi, however, Mary provides less evidence of how she actually evaluates the language provided in the performances through the lens of the descriptors and the scales. Here, Stormi's notes seemingly show more authority and document her decisions rather than focussing on documenting examples from the performances. It seems as if Mary tried to capture a lot of evidence and detail in her notes but was then overwhelmed when it came to taking decisions. This coincides with her comments on using the criteria (see Table 6.4), which signal excess cognitive load as Mary reported struggling with the number of descriptors in certain scales, keeping up with the performances as they unfolded (see comment on FLIN), and dividing her attention between different features.

Response to exit question. When asked about her perception of the rating process, Mary particularly focussed on how difficult she found it to apply the analytic rating scale: "because there are so many aspects to keep in mind at the same time." She admitted that she sometimes had to make a choice and could not consider all features of the performance at once: "sometimes you have to decide to shift your focus on one aspect and keep the others aside or on a back burner, because you simply can't consider each aspect simultaneously." The main problem for Mary is the level of detail in the rating scale which, on the one hand, she finds useful as a guide to know what features she is supposed to pay attention to, but which she also finds stressful as one can only hear each performance once.

Summary. Mary was a rater who obtained particularly low accuracy scores compared to the other participants. Her ratings were inaccurate as well as inconsistent, with quite a difference between her accuracy and severity for different criteria, and signs of erratic severity levels overall and in particular for the criteria RSL and ASL. Mary did not obtain notably high scores on any cognitive task and was below the average cohort performance in the Letters-Numbers task and the Trail Making task. Mary's overall results from the REI-40 indicated a preference for rational processing which also emerged from the decision-making style inventory. The long deliberation times are another noteworthy feature of her case.

The overall low accuracy and consistency of Mary's rating decisions appear to reflect that she found rating an overwhelming experience during the experiment. In her comments regarding the criteria, Mary seemed to struggle particularly with the language related criteria, ASL and RSL, and the level of detail required by the number of descriptors included in these dimensions. While the handwritten notes illustrate that Mary managed to consistently capture a lot of detail for each performance, few of her notes indicated a more global processing or judgement as was evident in some of Stormi's or Lexi's comments. Preferring the *rational* DMS, Mary appears to have taken meticulous notes in order to have a lot of information available for taking her decisions and considering the various aspects connected to each criterion. However, this need to process as much evidence as possible might have overwhelmed Mary as she stressed several times when responding to the exit question that she was struggling to focus and direct her attention during the rating process. It is possible that this was exacerbated even more by her tendency towards the *dependent* DMS, leading her to delay decisions as she could not look to or consult with others for her decisions. Thus, she deliberated on her notes for a longer

period even once the speaker had finished, while more successful raters seem to have approached rating more flexibly and independently as they integrated their observations as the performances progressed. Mary's case, therefore, seems to be shaped by an *overly* rational approach paired with a hesitancy to take independent decisions.

6.6 Case Study 4: Betty

When Betty participated in this study, she was 22 years old and one of the youngest raters ($M_{age} = 25.3$, $Mdn_{age} = 24$). She was in her sixth semester and less advanced in her studies than the average participant ($M_{semester} = 8.6$, $Mdn_{semester} = 7$).

Accuracy. Similar to Mary, Betty's ratings were less accurate than most participants' in this study. Measures of consensus or consistency with the reference scores were among the lowest of all raters (13.33 % exact agreement, $k = .14$, $r = .56$, $\tau_b = .47$). Only four raters obtained lower consistency metrics than Betty.

Betty's accuracy estimate (1.02 logits, $SE = 0.09$) from the MFRM RAM analysis is half a logit below the group. Similar to Mary, it is the criteria FLIN (1.19 logits, $SE = 0.21$), RSL (0.92 logits, $SE = 0.2$) and ASL (0.73 logits, $SE = 0.19$) where Betty's ratings were particularly inaccurate compared to the group (FLIN 1.8 logits, $SE = 0.24$, RSL 1.58 logits, $SE = 0.27$, ASL 1.65 logits, $SE = 0.23$). Betty's TA accuracy estimate (1.29 logits, $SE = .20$) was just slightly below the group mean (1.47 logits, $SE = 0.21$). The $MS_w = 0.60$ associated with Betty's ratings is indicative of either a reduced range or less risky decision making.

Severity. There were some noteworthy similarities and differences between Betty's and Mary's severity and fit estimates. Betty was even stricter than Mary in the

overall full model (ranked 3rd, 0.67 logits, $SE = 0.08$). For FLIN (0.81 logits, $SE = 0.19$), RSL (0.76 logits, $SE = 0.19$) and ASL (1.1 logits, $SE = 0.19$), her ratings were estimated among the harshest of the cohort (i.e., within one standard deviation of the maximum logit score). However, unlike Mary, Betty was also strict when it came to rating TA (0.71 logits, $SE = 0.17$).

In terms of rater fit, Betty's ratings showed a tendency of restriction and overfit. In the analysis based on the full model as well as the criterion-specific analysis of RSL and ASL, fit statistics were within the threshold considered useful for measurement (full model $MS_w = 0.77$, $t_w = -1.91$, RSL $MS_w = 0.83$, $t_w = -0.64$, ASL $MS_w = 0.76$, $t_w = -0.93$). For the criterion FLIN, the ratings were muted ($MS_w = 0.7$, $t_w = -1.14$), and for the criterion TA they were significantly restricted and lacking range ($MS_w = 0.49$, $t_w = -2.42$), indicating a tendency towards overfit.

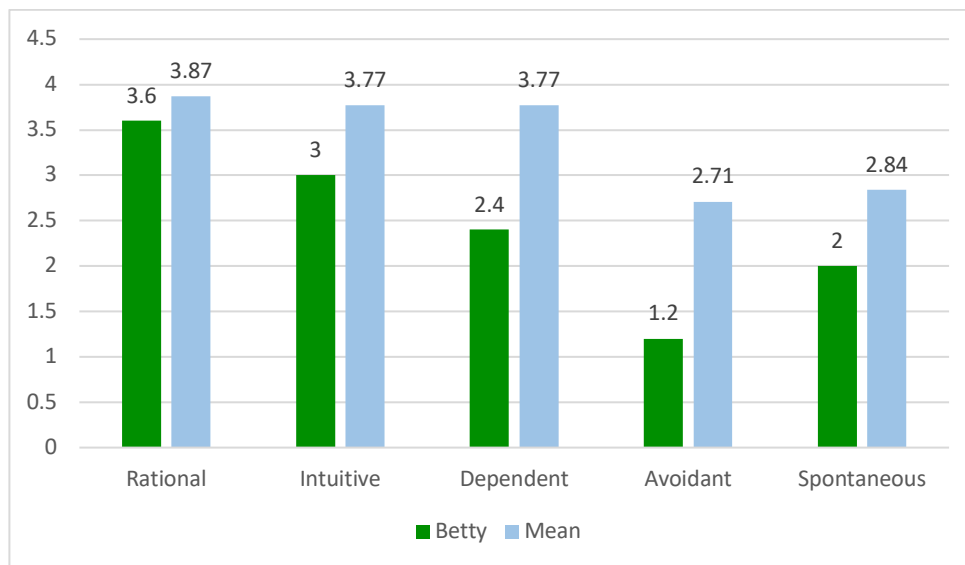
Cognitive variables. Betty's performance in the cognitive tasks was mostly below average. She ranked 26th in the Keep Track task (62.96, $M_{Keep\ Track} = 67.24$), 30th in the Stroop test (98, $M_{Stroop} = 148.53$) and 35th in the Trail Making task (49.98, $M_{Trail\ Making} = 67.77$). Betty did better on the Letters-Numbers task, where she ranked 18th (528, $M_{Letter-Numbers} = 472.95$) and her Digit Span score of 7 is right at the cohort's median performance.

Cognitive preference. Betty's responses to the REI-40 questionnaire indicate that she tends towards a balance between the two processing preferences. Overall and for both subscales, engagement and ability, Betty rated her preference for rational processing at a mean score of 3.1 which was slightly below the group means ($M_R = 3.66$, $M_{RA} = 3.56$, $M_{RE} = 3.77$). While her responses to the experiential ability subscale (3.70) were close to the group mean ($M_{EA} = 3.74$), her responses to the

experiential engagement scale (2.90) showed a stronger dispreference towards this subscale ($M_{EE} = 3.70$). Her combined experientiality score of 3.30 is also clearly below the group mean of 3.73. Betty's PSI score of 1, which is based on all responses to the REI-40, would identify her as a low experiential processor as Gunnell and Ceci's theory (2010) stipulates that one processing mode has the potential to override the other.

Decision-making style. Betty tended to choose lower scoring options for all DMS (see Figure 6.10). Within her profile, she preferred the *rational* DMS ($M = 3.6$) over the four other styles, making her the only selected case where the DMS scores do not reflect the results from the REI-40 and PSI. Her responses also indicated an aversion toward the *avoidant* DMS ($M = 1.2$).

Figure 6.10 Betty's DMS profile



Rating behaviour metrics. With an average deliberation time of 40.63 seconds, Betty was somewhat between Lexi, who was fast ($M_{DT} = 24.05$ s) and Mary who took longest of the four raters ($M_{DT} = 73.29$ s). However, she was still clearly below the cohort mean ($M_{DT} = 60.16$ s). Betty took longer than many until taking down

her first decision ($M_{TTFD} = 317.99$, ranked 29th). She was only one of three raters that did not revise any of their rating decisions.

Perception of rating. Betty's perception of the difficulty of each criterion remained quite stable over the two rating sessions. Her perception of TA and RSL (both $M = 3$) are clearly above the group means for these two criteria (both $M = 2.3$) and she felt that ASL was more difficult (2, $M_{difficulty\ ASL} = 2.37$). As far as the criterion FLIN is concerned, it is the only criterion where Betty's responses indicated some kind of change between the two rating sessions. Overall, however, her judgement of this criterion (2.5) agrees with that of her peers ($M_{difficulty\ FLIN} = 2.65$). Taking all of Betty's difficulty ratings together, she seemed to have found rating the individual criteria less challenging than any of the other three case studies and most of the other participants.

The two criteria that Betty perceived to be considerably easier to rate than most raters were TA and RSL (see Table 6.5). Concerning TA, Betty claimed that she found it easy to consolidate the performances with the actual descriptors in the scale. After Session 1, she reported that the speakers had clear instructions and that this made rating easier. After Session 2, Betty commented that she adapted her rating strategy ("did it differently"), however, the Likert-type items did not capture a change in her perception of the difficulty of this criterion. As the MFRM analyses for TA showed, Betty's rating decisions for this criterion were inaccurate, strict but also significantly restricted. One possible explanation for why she found rating easier is that she might have adopted a more superficial approach and restricted her observations to a quite limited interpretation of the construct.

Table 6.5 Betty's comments about the criteria

Criterion	Session 1	Session 2
	easy	easy
TA	Because the students are clearly told what they have to do and it makes it easier to assess it.	Did it differently than last time, now I am used to the format
	difficult	easy
FLIN	The most difficult was <i>strain on the listener</i> as there were frequently sections that were difficult to follow – how are you supposed to rate this? ⁸	I still think that fluency depends more on your gut feeling but the descriptors are somewhat helpful to justify your point of view
	easy	easy
RSL	You notice quickly whether someone has sufficient vocabulary to express themselves adequately	If you notice some specific things it is rather easy to do but can be difficult if you do not notice specific points
	difficult	difficult
ASL	Quite difficult because you notice a few mistakes but as you do you don't follow the rest [of the performance] – too focused on grammar	Because you may not always notice mistakes or are not sure whether or not it is accurate in the specific moment

As far as RSL was concerned, Betty's comment after the first round of rating appeared quite pragmatic: "you notice quickly whether someone has sufficient vocabulary". Her perspective was more nuanced after Session 2, as she highlighted that rating this criterion was easy if any striking or "specific" features come up but can be tricky when such clear characteristics are missing. While only mentioning lexis in Session 1, she broadens her comment to "specific things" in the Session 2

⁸ the statements for FLIN, RSL and ASL were translated into English. Betty originally wrote them in German. FLIN: „Am schwierigsten "strain on the listener" da es oft Teile gibt, die das Zuhören erschweren - wie bewertet man das dann.“

RSL: „merkt man schnell ob jemand genügend vokabular hat um sich passend auszudrücken“

ASL: „eher schwierig da man manche fehler bemerkt und dann aber dem weiteren nicht folgt - zu fokussiert auf grammatik“

which is more in line with the rating scale. However, given the accuracy of her ratings, her approach might be overly simplistic in that she appears to have defined a stopping or decision rule as soon as she notices “sufficient vocabulary”. As she points out herself, such features might be difficult to notice at times, particularly if her expectations are quite high as suggested by her severity.

Concerning FLIN, Betty’s perception shifted somewhat from finding it “difficult” after the first round to “easy” after the second. Betty struggled with the wording of the FLIN scale and the concept of “strain”. After rating more performances, she continued to struggle with the descriptors. While she found the descriptors in the scale “somewhat helpful to justify” rating decisions, she admits to relying on “gut-feeling”. Given the wording of her comments, Betty might not have managed to create a coherent and explicit understanding and representation of the criterion and the task of rating, leading to difficulties with matching descriptors to performances and a lack of differentiation between the ability levels.

The criterion ASL seems to have remained difficult for Betty throughout both rating sessions. One aspect she struggled with was that focusing on this criterion might inhibit noticing other features in the performances (“you don’t pay attention to the rest”). She also disapproved of the scale itself, but while Mary struggled with the level of detail in this criterion, Betty claimed that the scale focussed too much on grammar. After the second round, she still commented on the difficulty of noticing mistakes, but also highlighted that it might be hard to come to fast decisions about the appropriacy or accuracy of the language while rating the performance. It is noteworthy that there are several similarities between Mary’s and Betty’s comments about this criterion in that both criticise the scale, were not always sure

whether something they heard in a performance was actually a mistake or struggled with dividing their attention between accuracy and other features of the performance.

Notetaking in rating forms. Betty's notes on the performances P01 and P04 can be seen in Figure 6.11 and Figure 6.12, respectively. One striking feature of Betty's notetaking is that right from the beginning her notes were quite consistent and systematic on the surface level. However, they became considerably more detailed and specific in Session 2. While Betty often focussed on writing down the descriptors she applied at the beginning, there is much more detail including comments or quotes from the performances in Session 2.

As far as TA is concerned, Betty systematically documents her decision making. She writes down the descriptors she applied (e.g., "8.1, 8.2, 8.3") and supplements these with further notes; e.g., "brings in own opinion, relevant examples" or "gives good examples and arguments". Throughout Session 2 her notes evolve to include three categories for the three bullet points ("contrast", "family", and "school") with accompanying symbols.

The notes for the criterion FLIN most often contain the descriptors Betty applied to the performance and key words from the scale such as "fluent", "spontaneity", or "remarkable fluency". However, she also mentions elements such as the structure of the performance ("has a clear structure") and whether candidates had trouble when producing more language after prompting from the interlocutor ("has difficulties to talk about 2nd bullet point again"). The aspect of structure is not mentioned nor intended to be captured by the criterion FLIN and it is unclear why Betty seems to have chosen to include it in her decision making. The comments

regarding prompting by the interlocutor might help explain why Betty's ratings are harsher for this criterion as this only refers to one part of the performance that has to be produced entirely spontaneously and in reaction to the interlocutor's question. If Betty has given this aspect more weight than other raters, this might partly explain her strict ratings.

Figure 6.11 Betty's notes on P01 (rated 15th in Session 1)

Performance:	15		
TA	FLIN	RSL	ASL
8	7	8	6
Justification:			
TA: 8.1, 8.2, 8.3 → brings in own opinion, relevant examples			
FLIN: 8.1 very fluent start, 6.3.			
RSL: 8.1 - makes her point clear, good vocabulary → 8.2			
ASL: 6.4 - natural pronunciation but some mistakes 'deneked' → 'doncette'			
Rater training			
↳ "how less food" →			
↳ "purchases" → purchase			
↳ "avarnes" → awareness			

Figure 6.12 Betty's notes on P04 (rated 2nd in Session 2)

Performance:	2		
TA	FLIN	RSL	ASL
6	8	6	6
Justification:			
TA: 6.1, 6.2 - gives good examples + arguments			
FLIN: very fluent + good structured - 8.1			
RSL: 'food which... is not being eaten' → 6.3 / 'by having different...' → 6.4			
↳ 6.1 → sometimes she struggles finding the right words			
ASL: 6.3 - "good taste" - "food waste"			

For RSL, Betty mainly noted down the numbers of the descriptors in Session 1, with a few quotes from the performance or observations. The quotes often

document how a speaker produced various forms for one concept (“homeless people – people who are poor”), a particularly striking word choice (e.g., “benefit”, “reuse”, “cattle”) or comments such as “has good vocabulary” or “sometimes she can’t find the right word -> ‘there is a date that ...’”.

These features can also be observed in Betty’s notes regarding ASL. Here she would first mainly write down just the descriptors, and in Session 2 she would often add a few observations from the performances, with a focus on pronunciation (“auerness” vs. “awareness”, “pupil” vs. “people”) or correction of slips and mistakes. Considering that Betty criticised the scale for emphasizing grammar, she takes hardly any notes documenting decision making in relation to grammatical accuracy; decisions for this descriptor (.2) are usually missing while decisions for the other descriptors of this criterion (.1, .3 or .4) are documented for most speakers. There is, therefore, little evidence of how she might have interpreted this descriptor and or how she might have weighed her observations.

Another aspect that can be explored through her notes is Betty’s particularly severe interpretation and application of the scale descriptors. For instance, her notes on FLIN reveal that she observed a speaker to be “fluent, few longer pauses” or even “very fluent”, but still decided to choose bands 6 or 8, respectively. Similarly, she highlights the range of a candidate by noting down examples of use of passives (“food which is not being eaten”) and gerunds (“by having different...”) (see Figure 6.12), but ends up awarding band 6 to a performance that received a fair score of 8 for this criterion. It is also striking how Betty weighs the descriptors, when her notes read 6.3 and 8.1 in RSL and she eventually awarded band 6 for this criterion overall.

Judging only from her notes, it will remain unclear as to why Betty's ratings were so severe and inaccurate. One explanation could be a lack of differentiation. There is evidence for differentiation in her notes, but the features Betty captures and the descriptors she notes down did not translate into differentiated scores. Thus, speakers that appeared to be quite different from each other in her notes ended up with similar bands. It seems as if other factors than those that were documented significantly impacted on the rating process. Another explanation could be that Betty misinterpreted the scale and what a band such as 6, which is essentially the minimum pass grade, actually means. After all, to someone who is not familiar with the scale, a band 6 might appear to be an achievement on a scale of 0 to 10. Finally, it might also be the case that Betty based her decisions on a distorted or reduced representation of the construct as can be seen in her notes on FLIN and ASL. As a result, Betty's decisions rest on fewer and isolated observations about the performance.

Response to exit question. Betty mentioned several aspects about the experiment that she found challenging. First, she admitted that she did not quite know how long the rating session would be and that this worried her while rating the performances.⁹ Second, Betty felt that she was not entirely consistent and experienced rating as quite strenuous. She mentioned that knowing some of the speakers was challenging

⁹ Betty received the same instructions as all participants. The overall format of the rating sessions were on the cover page of the rating form (see Appendix F **Error! Reference source not found.**). The rating forms also indicated to the raters how far they had already progressed in the experiment. However, this concern might have impacted Betty's ratings in the Session 1 when everything was still less familiar.

for her as she tried to stay neutral in her ratings.¹⁰ Betty reported that starting to use the rating scale was not “all that hard” but concedes that,

“if you familiarize yourself more with [the scale] in the beginning it would be a lot better.”

Just a few sentences later, Mary admits that she felt that getting to know the scale and starting to work with the performances was quite “overwhelming”. Despite being just a short interview and an open question, Betty’s comments do not appear consistent, and even contradictory, as she explains that she found applying the criteria easy on the one hand, but also explains that it was a strenuous activity for her.

Summary. Betty was the second inaccurate rater included in the case studies. The exact agreement of her scores was extremely low and they also showed little consistency with the rating decisions of other raters. While Mary’s ratings tended to be erratic and less predictable, Betty’s severe ratings appeared restrained and tended towards overfit. Overall, her performance on the cognitive tasks was weak in comparison to the cohort. Her responses to the REI-40 revealed no pronounced preference for either, rational or experiential mode, but the PSI indicates a slightly stronger orientation towards experiential processing. The GDMSI showed a preference for the *rational* DMS over the other styles. Betty did not deliberate her rating decisions overly long but held back with entering her first decision longer than many of the other raters. Once she had entered a decision, however, she did

¹⁰ Three of the 30 performances included in the experiment were performances from university students.

not change it. Her self-report responses concerning the difficulty of rating but also her comments from the interview showed that she seemed to find rating easier than many. Betty's handwritten notes revealed details about her strict and often less differentiated rating decisions.

Betty's case presents an intriguing combination of features. On the one hand, she was systematic in her notes which might fit a more *rational* DMS and her high levels of confidence and late first decisions could explain why she did not revise any of her rating decisions. However, the fact that she was quite confident about her rating decision and did not find rating difficult does not fully align with her comments from the interview, where she listed various challenges of the rating process – e.g., the scales, lack of familiarisation, or difficulty of rating some speakers. Furthermore, Betty's notes were often sparse, offered little to go on once the performance had finished, and were found to be at odds with her final decisions (e.g., how she weighed the descriptors). Betty's conceptualisation of successful rating may be that she was required to be systematic, thorough, but also strict with her decisions. Betty did not feel that rating was particularly difficult and unconscious processes might have remained unchecked by metacognitive strategies. These observations suggest that Betty's rating behaviour might have been shaped by surface-level processing. This is also supported by the fact that she chose lower values in the self-report questionnaires, entailing that she neither enjoys challenging cognitive tasks nor believes in her ability to engage in either cognitive mode. Further, unlike Lexi, the strain of rating was not successfully mediated by metacognitive strategies.

6.7 Between-case Comparisons

The goal of the case studies was to investigate what features of rater cognition may have contributed to the differences in accuracy of four raters. Four raters were selected to represent extreme cases as far as accuracy was concerned. DMS profiles as well as their handwritten notes were also considered. Based on the data collated and presented in this chapter, more detail can be added to each of the four rater profiles (see Table 6.6).

Table 6.6 Detailed profiles of Mary, Betty, Stormi and Lexi

Low accuracy	High accuracy
Mary – overly rational and dependent <ul style="list-style-type: none"> erratic rational and dependent DMS rational PSI (-10) rational, detail-oriented, slow processing rigid and indecisive overwhelmed by detail of both, performances and scales 	Stormi – balanced and flexible <ul style="list-style-type: none"> balanced: rational & intuitive DMS experiential PSI (3) quite fast decisions & frequent revisions structured & flexible independent
Betty – surface-level processing <ul style="list-style-type: none"> consistent rating rational DMS experiential PSI (1) overly confident and no revisions structured notes surface-level processing data less coherent – possibly stronger influence of unconscious processes 	Lexi – intuitive and flexible <ul style="list-style-type: none"> intuitive & spontaneous DMS experiential PSI (13) swift decisions but still structured notes open to revisions & appears to self-regulate formulates decision rules based on scale, which do not override engagement fatigue – uneven cognitive performance

While Stormi and Lexi's cases have shown that a structured approach, a degree of rigour and meta-cognitive regulation may have been helpful to mediate possible

pitfalls of intuitive decision making, Mary's case demonstrated that an overly detail-oriented approach may have adverse effects in that it increased cognitive load and impacted rating performance. Furthermore, the highly accurate raters displayed an in-depth engagement with the descriptors leading to more encompassing judgments that were nonetheless in line with the construct. The less accurate raters, instead, seemed to focus more on noticing key features. However, conceiving rating to be about thorough documentation did not contribute to accurate rating that was aligned with the scale and the construct. Study 2 presented in the previous chapter did not produce many statistically significant results. The cases included in this chapter, however, point towards some possible interactions between cognitive capacity and decision-making styles that might have been too complex (due to mediation effects) or too weak (in terms of effect size) to be captured with the sample size included in this study.

7 Summary and Discussion

7.1 Outline

In this chapter, the results from the three studies presented in Chapters 4, 5 and 6 will be summarized and discussed more extensively before moving on to a wider discussion about the conceptualisation of the rating processes when assessing speaking. It will recapitulate the aims, methodology and research questions in Section 7.2. Sections 7.3 to 7.5 are dedicated to discussing the three studies separately. Section 7.6 will discuss the findings more broadly by focussing on how they contribute to our understanding of the role of the raters, rating scales, and rating process in the context of assessing speaking. Section 7.7 will examine how findings from this study may be integrated into an argument regarding the evaluation inference. The practical implications of this study will be discussed in Section 7.8 and Section 7.9 will talk about the merit of some of the research methods employed.

7.2 Introduction

The aim of this study was to explore the nature of the rating process of spoken language, particularly as it is experienced by novice raters. A mixed-methods study was conducted to address this aim. Thirty-nine participants received a brief introduction to the Austrian speaking scale which is currently used for the B2-level Matura, with a chance to rate and discuss each criterion, before rating 30 video-recorded performances via an online survey tool. While raters entered their rating decisions into the online rating environment, embedded JavaScript was used to

record time stamps for each rating behaviour initiated by a mouse click: the start and end of rating a performance as well as each rating decision.

Raters filled out two validated self-report questionnaires about their preferred cognitive processing mode and decision-making style prior to the rating procedure. In addition to the actual scores, data was collected throughout the experiment by Likert-type self-report items embedded within the online rating procedure and via handwritten notes that raters provided for each performance. After the rating sessions, raters were individually tested on their cognitive capacity before completing their participation by responding to a set of exit questions.

The main research focus of exploring the nature of rating spoken language was operationalized via three main research questions:

1. How does a group of novice raters use the analytic rating scale currently in use for assessing speaking performances in the Austrian Matura examination?
2. Are the rating quality and rating behaviour metrics produced by the novice raters related to cognitive attributes, preferred processing mode or decision-making styles?
3. To what extent do case studies of accurate and inaccurate raters reveal differences in rating behaviour and influences on rater behaviour?

Various analytical methods were used to investigate the data derived from the rating sessions, the questionnaires, the cognitive tests, and the exit questions. The scores provided by the raters were analysed through common CTT methods and two MFRM approaches, the Rating Scale Model (RSM) and Rater Accuracy Model

(RAM). The specific features investigated were inter-rater agreement and consistency (CTT), intra-rater consistency (RSM fit analysis), rater severity (RSM), and rater accuracy (RAM). Other aspects related to scale functioning, for instance, criterion measurement and bias were also considered.

The timestamp data was used to create three metrics of rating behaviour: 1) deliberation time, 2) time to first decision and 3) number of revisions. The various sources of self-report data (questionnaire data, perception data during and after rating) were analysed quantitatively and used to create variables for cognitive capacity, cognitive preferences, and decision-making style. Data from open items about rating scale use were explored through thematic analysis. Finally, the comments from the exit question and handwritten notes were used to supplement selected cases in a set of case studies.

7.3 Study 1: Rating Quality and Scale Use

7.3.1 Summary

Study 1 focussed on addressing the research question: How does a group of novice raters use the analytic rating scale currently in use for assessing speaking performances in the Austrian Matura examination? Drawing on Knoch and Chapelle's (2017) recommendations regarding the evaluation inference, the consistency and accuracy of ratings as well as bias patterns and perception of scale use were investigated.

The ratings were found to be of highly varying quality with considerable inter-individual variation. Some novice raters were successful in reaching acceptable

indices of inter-rater consistency, intra-rater consistency and accuracy, while others struggled. As a group, the raters reached more agreement with the reference scores than between themselves. The analysis of both MFRM models, RSM and RAM, revealed the heterogeneous nature of the rating decisions and provided evidence of model underfit and bias for several elements of the MFRM models, i.e., for several raters, performances and criteria. Raters also varied in their perception of difficulty of rating in general and of each criterion. Rater perception of criterion difficulty triangulated the results from the MFRM analysis. The thematic analysis highlighted the specific challenges attached to the application of each criterion.

While raters were, broadly speaking, reasonably successful with applying the rating scale, the criterion TA frequently emerged as the most problematic. This was confirmed by all three analyses (MFRM RSM & RAM, and perception data).

7.3.2 Severity – Discussion

While it is evident that the raters varied considerably in terms of their consensus ($M_{\%agreement} = 30.49\%$, $M_K = .33$) and consistency ($M_{\text{tau-b}} = .49$), their performance compares to that of raters in other studies. The raters included in Davis' study (2008, 2015), reached an exact agreement of 38.5% and weighted kappa of .43 towards the end of the experiment after rating more than 400 performances. Their mean pairwise inter-rater consistency reached .67. However, Davis' group of raters were highly trained and experienced EFL teachers from the outset while the group of raters included in the current study had not completed their training and had only little classroom experience. Furthermore, the scores in Davis' study were based on holistic decisions rather than rounded analytic rating decisions. Thus, from a

psychometric perspective this could also explain why there is more discrepancy in the current study.

The MFRM-based estimates of rater severity in the current study are similar to findings in other rater reliability studies (Eckes, 2015; Davis, 2008, 2015; Schaefer, 2008). Consistency in ratings was comparable to that of Davis (2012), whose sample consisted of highly experienced teachers, or displayed even less variation than found in other studies (e.g., Wang & Luo, 2019). There was a slightly wider range of 3.25 logits in Schaefer's study which included 40 novice raters. These results, therefore, are in line with expectations from previous research. They also provide further support for the claims that raters can never fully agree in terms of severity (McNamara, 1996) and that consistency is not likely a feature that distinguishes raters at different levels of expertise (Lim, 2011).

In previous studies, rater fit is frequently considered an important indication of how well raters adapt to the challenging task of rating language performance (Eckes, 2012; Lumley, 2005; J. Zhang, 2016). The proportion of misfitting ($N = 7$) and overfitting ($N = 5$) raters in a group of 39 individuals appears larger than in many rater reliability studies (e.g., Schaefer, 2008; Yan, 2014). To some extent, this was to be expected given that the thresholds were set narrower than, for example, in Yan's study, where upper and lower limits were defined quite generously between 0.5 and 1.5. One goal of this analysis was to identify and investigate misfitting and overfitting raters, thus choosing narrower thresholds was in the interest of highlighting variability.

Nonetheless, intra-rater consistency was poorer in this study compared to other rater reliability studies even after reviewing thresholds. One explanation could be that

the raters were novices. However, various rater studies failed to establish a link between rater expertise and rater fit patterns (Davis, 2015; Kim, 2015; Lim, 2011; Schaefer, 2008). The rater consistency in Davis' cohort of 20 raters shifted from two misfitting and two overfitting raters in session one (before training) to four overfitting raters in the last session, but when comparing the fit statistics of the entire cohort over the four data collection points, Davis still concluded that fit might not be impacted by training. In the context of writing, the 40 novice raters in Schaefer's study were quite consistent with only three misfitting and two overfitting raters. Barkaoui (2011) on the other hand, found that novice raters tended towards misfit in both, analytic and holistic rating while experienced raters were more likely to be overfitting. In absence of expectations as far as intra-rater consistency is concerned, we cannot say with certainty whether the lower consistency in the current data was due to the fact that raters were novices to rating and had no teaching experience, the nature of the analytic rating scale, or the complexity of the construct.

One factor that is likely to have contributed to lower intra-rater consistency is the varying difficulty and fit of the criterion facet. The criterion TA was significantly easier than ASL or RSL, displayed misfit and was associated with a majority of residuals. There were considerable interactions between different facets and a high number of unexpected observations associated with the TA element. However, overall, the number of statistically significant residuals ($t > 2$; $N = 212$, 4.53%; $t > 3$; $N = 34$, 0.72% of 4,680 observations) was within acceptable parameters. As far as differing severity levels is concerned, the finding that language accuracy or language resources are rated more harshly is consistent with other studies investigating the assessment of speaking (e.g., McNamara, 1996; Youn, 2018) as

well as writing (e.g., Ballard, 2017; Eckes, 2015; Lumley, 2005; Schaefer, 2008; Wang & Engelhard, 2017). As far as fit is concerned, McNamara (1996), Eckes (2016) as well as Wang and Luo (2019), and Wang and Engelhard (2017) also found lower and more dependent estimates for the expression-related criterion and noisier estimates for the content criterion. A bias toward the content category was, for example, also detected by Wang and Luo (2019) who concluded that “the training program was not effective in clarifying its meaning” (p. 106). None of these studies attempt to explain the reasons behind this effect.

While the model fit for all five RSM models investigated was acceptable, the results raise questions about the latent variable modelled via the MFRM approach. The element TA displayed local misfitting patterns in the full model, but also the weakest global fit in the criterion-specific analyses. Furthermore, severity in TA produced the weakest correlation with the full model severity measure. These observations highlight that it is unlikely that all four criteria included in the rating scale map onto the same latent variable in a comparable or equal manner. While it was still possible to successfully fit the rating scores to the model, this finding emphasizes the breadth of the construct that raters were asked to assess via the rating scale and the considerable noise contributed by the criterion.

7.3.3 Accuracy – Discussion

Based on the CTT analysis of the holistic scores, there was a considerably larger degree of consensus and consistency of the novice raters with the reference scores than with each other. Consensus indices in this cohort ($M_{\%agreement} = 35.56\%$; $M_{\kappa} = .38$) are similar to those reported by Davis (2008, 2015) prior to training ($M_{\%agreement} = 33.9\%$; $M_{\kappa} = .35$).

The RAM MFRM analysis generally produced a better global model fit than the RSM analysis. This might be due to the narrower range of scores and more observations per score category which are a direct result of converting the ratings provided by the raters into accuracy scores. Consequently, there were fewer residuals and underfitting raters, as well as a narrower range of accuracy estimates in the RAM approach.

The criterion estimates from the RAM analysis indicated statistically significant differences in how accurately a criterion was rated. The criterion TA was the easiest in terms of severity (in the RSM), but the most difficult to rate accurately. This is consistent with what Engelhard et al. (2018) found in their study of a two-criterion analytic rating scale in a writing assessment. Again, similar to the results based on the RSM and in Engelhard et al. (2018), the TA facet also produced slightly more unexpected observations than the other facets. This was to be expected because when the ratings related to a criterion are more erratic, as was established to be the case for TA in the RSM, the chances of being further removed from the reference score also increase, leading to unexpected low accuracy scores.

An exploratory bias analysis revealed that high proficiency speakers towards the upper end of the ability scale seemed to attract significantly more unexpected inaccurate or accurate ratings. This finding is in line with other rater reliability studies (Schaefer, 2008; Wang & Luo, 2019). The fact that the only other noteworthy bias concerned a speaker at the low-end of the ability spectrum also matches Kondo-Brown's results (2002).

Any discussion of rater accuracy, particularly the results in the current study that are based on the RAM MFRM model, is limited by the fact that studies usually do

not include an external criterion such as reference scores and report on these. This might be due to the fact that intra-rater and inter-rater consistency (within a cohort of raters) is emphasized or has been investigated more often than rater accuracy. Through techniques such as anchoring there is often less of a practical need in operational testing to investigate rater accuracy via comparison to reference scores. However, this study has demonstrated the usefulness of considering both consistency and accuracy in tandem in exploring the nature of novice rater behaviours.

7.3.4 Perception Data – Discussion

The perception data revealed that more rater confidence coincided with lower perceived difficulty of rating. Interestingly, there was no statistically significant correlation between confidence levels and perceived difficulty and rating quality.

The finding that raters' perception varied between criteria is consistent with other studies that investigated rater confidence in the context of large-scale test development (Pearson Test of English Academic; Ackermann & Kennedy, 2010) and score validation (IELTS Speaking Test, Brown & Taylor, 2006; Cambridge English Language Assessment, Gilbert & Staub, 2014). A survey of Swiss examiners of the Cambridge English Assessment Suite found that even for experienced assessors, factors such as multi-tasking as assessor and examiner, difficult-to-rate performances and too brief descriptors can pose a challenge (Gilbert & Staub, 2014). It is, therefore, not surprising that the novice raters struggled with various features of the scale. It is, however, noteworthy, that the perception data converged with findings from the MFRM (RAM) analysis. As a group, the raters successfully predicted the relative difficulty of rating the criteria

accurately. The RAM analysis ranked the criteria from easiest ($\text{logit}_{\text{FLIN}} = 0.12$) to hardest ($\text{logit}_{\text{TA}} = -0.21$) which matches the overall perception of these two criteria.

The thematic analysis of the justifications illustrated the various reasons why certain criteria may be more difficult to rate than others. For some raters the level of detail in the RSL and ASL scale was a particular challenge as there is a larger number of descriptors and more features raters can look out for, while others felt that FLIN was easier to rate as there were fewer descriptors. This resonates with frequently discussed concerns about descriptor wordings. On the one hand, striking a balance between detail of description and usability is a great challenge for scale development. Streamlined descriptors that are fast to use and access in an exam situation (i.e., user-oriented according to Alderson's 1991 classification) may pose challenges to the raters in the face of a broad range of performances they may encounter (see Gilbert & Staub's, 2014). On the other hand, detailed descriptors offer raters more scaffolding, at the expense of having to deal with more descriptors and to take more weighing decisions.

7.3.5 Factors Influencing Rating Quality

Rating quality in this study may have been impacted by several factors. A first major aspect likely to contribute to more diverse ratings is to be found in the performances included in the study. In many other reliability studies (e.g., Davis, 2008, 2015; Lim, 2011; J. Zhang, 2016) the ratings are provided on selected recordings of real test performances. However, video-recording a live examination was not an option for several reasons the most relevant of which was the fact that Austrian students are not tested on one standardized prompt, but many teacher-generated prompts.

Furthermore, the ratings in this study are likely shaped by the test taker experience of the speakers. While recordings from actual examinations might include speakers that are prepared or trained to perform well on these speaking tasks, some of the speakers in the recordings used for this study had to be prompted to reach the five minutes of speaking time required in the test specifications. Whenever the interlocutor intervened, the speakers needed time to react to the follow-up questions, leading to pausing and less fluent speech as speakers tried to come up with further talking points. This might have contributed to the discrepant ratings for the TA criterion and was also commented on by raters as some found it more difficult to rate such performances.

The length of the performances as such might be another feature leading to lower consistency and accuracy, particularly when novice raters are involved. With about five to six minutes of speaking time, performances were quite long relative to some other speaking proficiency tests. Longer performances have the advantage that there is more language to assess and even if raters get distracted by certain features they still have time to bounce back and gather more evidence. This could be particularly useful as the examination is not recorded and must be assessed in real time. However, the length of the performances in this study could also prove problematic for consistent scoring for several reasons. First, the chances for the speakers to produce uneven features within one performance are higher as they might get tired throughout the performance, run out of ideas previously prepared in the short preparation time or, on the other hand, warm up to the task and speak more fluently in the second half of the performance. This greater variation within and between the performances increases the chances for the raters to produce less consistent and accurate ratings as they have to deal with a larger number of cues to relate to the

descriptors in the scale and weigh local brilliant moments or mistakes in comparison to the whole performance. The longer performances also demand longer periods of concentration on part of the raters. They may also boost rater effects in that certain rater tendencies may be amplified as the performance progresses. If raters take a harsher stance, they might end up tallying more and more evidence to support their harsh judgement about a speaker's weaknesses and this may be the other way around for very lenient raters. It might require metacognitive regulation and reflexive thinking to balance out the large number of cues and observations.

The wording of the scale or training provided could also have contributed to rater variability observed in the data. The one finding that stands out in terms of the rating scale is the differential functioning of the content criterion TA. On the one hand it was the easiest criterion to score well on based on the RSM-based estimates. However, it was also the criterion that raters rated the least accurately. The question is whether more training or different wording of the rating scale is likely to improve this aspect given that the scores and estimates behaved similarly to other rater reliability studies. Having an exemplar that displays interlocutor intervention and how to deal with it when assessing TA could possibly have helped improve the ratings for the purposes of the study. On the other hand, it is a less likely scenario to occur in real live tests where students have longer preparation time and are keen to keep talking for as long as possible. As this finding reoccurs in other studies, it raises questions about the operationalizability of the content dimension in speaking assessments. Many studies, including this one, focus on large-scale speaking assessments (Fan & Yan, 2020) which tend to rely on general descriptions of task

achievement which in turn may be more difficult to match to many different tasks and student performances (Fulcher, 2003).

While rating quality was certainly problematic compared to the performance of raters in other reliability studies, comparing these findings with Davis' (2008, 2015) data might allow for a more encouraging interpretation. Davis' sample consisted of experienced teachers who are likely to have heard a lot of student talk in the target language and who are bound to have at least some experience in assessing language performance in other contexts. The analysis in the current study shows properties that are similar to the data in Davis' pre-training sample. One way to interpret this finding is that the novice raters in the current study were able to make up for their lack of classroom experience at least somewhat by receiving a minimal training and rating 30 performances. Furthermore, Davis' ratings are based on true holistic scores while the data for the CTT based analyses was aggregated and averaged arithmetically. Had raters been asked to provide an overall single band for each performance, the final score they would settle on might have been different to an arithmetic mean depending on how they consciously or subconsciously weighted the four criteria.

Finally, the sample of raters might have contributed to the observed rater variability. While care was taken to narrow the range of experience and expertise in the recruited participants there might be underlying factors impacting rating quality. For instance, some raters might have more experience in rating or teaching English than others if they have opted to work as tutors in one of the many tutoring institutes or have been able to involve themselves a lot in the assessment during their practical work at the schools. Differing levels of English competence might also play a role.

While all participants have at least a minimum level of B2 and, according to the university curriculum should be on their way to reach the C2 level, it was not possible to establish with acceptable precision the current English proficiency of the raters.

7.4 Study 2: Exploring the Role of Cognitive and Psychological Attributes in Rater Cognition

7.4.1 Summary

The previous discussion sought to explain the observed variability in rating quality. The aims of Study 2 were to investigate whether a set of specific and yet under-explored influences might contribute to rater cognition and rater behaviour metrics. These were the raters' cognitive capacity as measured by cognitive tests (Stroop, Letters-Numbers, Digit Span, Keep Track task, and Trail Making Task), their preferred processing mode as measured by the Rational-Experiential Inventory (REI-40), and their preferred decision-making style as operationalized in the General Decision-Making Style Inventory (GDMSI). Given the relative lack of experience among the novice raters in this study, and the fact that they received only minimal training, it was hypothesized that the influence of individual differences in cognition and decision-making would potentially have a stronger effect on rating behaviour. A series of correlational analyses was conducted. The associations detected in the data revealed that neither cognitive variables nor processing preferences were strong predictors of rating outcomes. The *intuitive* and *avoidant* DMS, however, were found to predict severity in RSL and ASL, as well as accuracy in RSL.

7.4.2 Cognitive Variables – Discussion

Overall, the cognitive tasks included in this study had no significant predictive power as far as rating quality and rating behaviour metrics were concerned. There were only few correlations above or below $\pm .30$. The Stroop task, measuring inhibition, displayed one correlation above .3 with deliberation time ($r_s = .37$) and the Keep Track task, measuring the ability to update information stored in the working memory correlated negatively with the number of revisions ($r_s = -.37$). Overall and in relative terms, the Stroop task produced the most consistent and strongest correlational patterns while the Digit Span task performed the weakest out of the five tasks included in the test battery. Any interpretation of the data would be speculative at this point and in light of the overall weakly expressed correlational patterns the results must be considered inconclusive.

Apart from the underlying complexity of the construct under investigation, there are several aspects which are likely to have shaped the results. First, the tasks may have failed to properly discriminate between the participants, or, on the other hand, the participants may have been quite similar with respect to the latent variable measured by the tasks. The measures of dispersion for the cognitive tasks ranged between 15% (Digit Span) to 52% (Stroop). Particularly the Digit Span, Keep Track and Trail Making underperformed in discriminating different ability levels. The participants may also have been more homogeneous in their WM abilities. All of them were pursuing academic studies and thus were already likely to be similar due to the selective nature of long school careers before even beginning with their studies. What is more, learner studies including WM measures tend to correlate WM-indices with fine-grained and highly sensitive behavioural data such as eye-

gaze metrics or keystroke logging metrics (e.g., Révész et al., 2017). As learner populations are accessible in larger numbers than rater populations, such studies were able to work with larger samples which is conducive to producing more variation and stronger effects along the variables included in the study (e.g., Indrarathne & Kormos, 2017; Kormos & Sáfár, 2008).

While it is unlikely that features of WM or attention play no role in the quality of rating decisions and rating behaviours, the current study could not detect significant effects of cognitive attributes on rating outcomes. However, this finding is similar to Isaacs and Trofimovich's findings (2010) who were also unable to detect significant effects of attentional direction and rating of spoken language. It seems likely that other conscious and subconscious processes, and their interplay, mediate the effects of phonological short-term memory and flexibility when rating spoken performance rendering the effects of WM alone to be trivial with respect to rating quality.

7.4.3 Preferred Cognitive Processing Mode – Discussion

The cohort of raters included in this study showed a preference for more experiential decision making compared to, for example, the sample of student pharmacists included in McLaughlin et al. (2014). This could be explained by the large proportion of women in the current sample ($n = 35$ out of 39) as women have been found to display a slightly stronger preference for experiential processing across several studies and populations (Pacini & Epstein, 1999). However, there is evidence that these gender differences may not be stable or have quite small effect sizes (Epstein et al., 1996; Shirzadifard et al., 2018). An alternative explanation might be that students interested in becoming language teachers generally display

a tendency towards experiential processing or at least a balance between the two processing modes compared to other student samples.

The preferred processing mode as operationalized in the REI-40 (Pacini & Epstein, 1999) was found to have no strong effect on rating quality and rating behaviour. There were no correlation coefficients above $\pm .2$ between the two overall scales, rationality and experientiality, and any of the rating quality measures or rater behaviour metrics. The hypothesis that a preference for rational processing or engagement might generally have a positive effect on rating quality, as it was found to have in other decision-making contexts (e.g., Gunnell & Ceci, 2010; Fletcher et al., 2012), was not supported by these findings.

While the REI-40 has been used successfully to measure individual differences in particular populations (e.g., student pharmacists in McLaughlin et al., 2014), to explore the connections between personality traits and behaviour (e.g., Big Five and REI-40 in Witteman et al., 2009) and to predict real-life outcomes (e.g., academic success in Shirzadifard et al., 2018) it was also found to underperform as a predictor in relation to specific reasoning tasks or cognitive tasks (Sánchez et al., 2012). After all, the REI-40 rests on the assumption that individuals are capable of self-reporting their ability and disposition in relation to certain processing modes. This may not reflect, however, the true ability or actual intensity of engagement with either cognitive system (e.g., Sánchez et al., 2012). This might partly explain the weak correlations between the REI-40 and rating quality. Another possible explanation could be that success in rating does not necessarily rely on one processing mode overriding the other as was suggested by Gunnell and Ceci (2010) when they developed the PSI (Processing Style Influence) in the context of juror decision

making. Instead, being able to switch or balance these two modes in reaction to the performance as it unravels in real time and directing our attention to fulfil certain goals might be more helpful to achieve consistent and accurate ratings (see also Phillips et al., 2015; or Fletcher et al., 2012).

7.4.4 Decision-making Styles – Discussion

Several subscales of the GDMS inventory predicted rater severity and accuracy. A higher preference for the *intuitive* DMS was associated with lower severity in the criterion RSL ($r_s = -.48$). The *avoidant* DMS correlated negatively with severity on RSL ($r_s = -.51$), severity on ASL ($r_s = -.50$) and severity on the full model ($r_s = -.52$). Furthermore, the *avoidant* DMS also correlated positively with accuracy on RSL ($r_s = .53$).

There are a few correlations of +/- .30 or higher that did not produce statistically significant results. Such patterns can be found between the *intuitive* DMS and all severity measures except TA (r_s between -.32 to -.48), the *dependent* DMS and all severity measures except FLIN (r_s between -.33 to -.39), and the *avoidant* DMS and all severity measures (r_s between -.38 to -.52). When looking at accuracy, correlations of about +/- .30 and higher can, again, be found between the *intuitive*, *dependent*, and *avoidant* DMS and accuracy on RSL and ASL (r_s between .29 to .53).

While the *intuitive* DMS has a comparatively firm theoretical footing in line with the cognitive-experiential self-theory (e.g., Thunholm, 2004), it is less clear what constitutes the *avoidant* DMS in addition to a basic reluctance to take decisions. Studies with large samples were able to confirm the five-factor structure of the

GDMSI (Loo, 2000; Thunholm, 2004; Spicer & Sadler-Smith, 2005; Gambetti et al., 2008), but it is still not fully understood what decision styles *are* in relation to different decision-making tasks (Berisha et al., 2018).

The results of Study 2 indicate that there is variation in what kind of DMS raters might prefer, but they are unable to explain *why* or *how* either style contributed to rating quality. An attempt at explaining the findings in the current study will have to remain tentative and draw on the behaviour of these variables in other studies. Furthermore, the styles are not mutually exclusive and were found to correlate positively in this sample as well as in numerous other studies (*intuitive* x *dependent* $r_s = .23$, ns; *intuitive* x *avoidant* $r_s = .19$, ns; *avoidant* x *dependent* $r_s = .53$, $p < .01$).

In the current set of results, the *intuitive* DMS correlated significantly positively with lower severity in the criterion RSL. The *intuitive* DMS was also found to correlate negatively with the *rational* DMS ($r_s = -.30$), adding to a range of erratic patterns in various other studies with correlation coefficients ranging from $-.21$ (ns, Thunholm, 2005) to $.16$ ($p < .01$, Delaney et al., 2015). The effect of the *intuitive* DMS when using an analytic scale as compared to the *rational* DMS on severity may at first appear counterintuitive. After all, faster decision making is frequently problematized as relying on heuristics and being prone to error in Kahneman's view (2003; Kahneman & Frederick, 2002) and a preference for slower, rational decision-making did not predict successful rating. However, faster intuitive processing has been found to be beneficial in tasks that are complex, rely on "subjective and incomplete information" (Akinci & Sadler-Smith, 2013) and are performed under time pressure where a purely rational approach might not be appropriate (Spicer & Sadler-Smith, 2005). Thus, these findings might shed light

onto the nature of the process of rating speaking in real time – at least for the context of this experiment – in that raters tending towards more intuitive processing were better able to balance the cognitive demands of rating the criterion RSL. The *intuitive* DMS might contribute positively to the decision outcome in that raters are less overwhelmed by the number of descriptors and endorse a more holistic or configurational approach. As noted elsewhere, the RSL and ASL scales have many descriptors, forcing raters to weigh many observations. Thus, being able to balance out the analytic observations with a more global impression (or ‘gut feeling’) towards the performance might have helped raters to emerge as less severe and more accurate in this study.

The reasons behind the success of the *avoidant* DMS might be slightly different. In this study, higher *avoidant* scores coincided with lower *rational* scores ($r_s = -.20$) and higher *dependent* score ($r_s = .53$), which was consistent with many studies that included large and varied samples, e.g., Delaney et al., 2015; Spicer & Sadler-Smith, 2005; Loo, 2000; Thunholm, 2004. The *avoidant* DMS, however, did not correlate significantly with any of the rater behaviour metrics (deliberation time, time to first decision or revisions). In other research, the *avoidant* DMS was found to correlate positively with traits such as hesitancy (Scott & Bruce, 1995), or indecisiveness and negatively with measures of decision success (Decision Outcome Inventory, or Adult Decision-Making Competence; see Bruine de Bruin et al., 2007). In a study investigating the decision-making of parents concerning school choice, Galotti and Tinkelenberg (2009) discovered that the *avoidant* DMS correlated with a larger number of resources consulted and criteria considered. Therefore, a need to gather more information might have benefitted the *avoidant* DMS. The *avoidant* DMS might also have affected rating quality in that raters could

be less likely to choose the more extreme scores which puts them less at risk for being exposed as overly severe/lenient or inaccurate. However, neither Baker's (2012) study nor this study could confirm a central tendency for avoidant raters. The *dependent* DMS, which did not produce significant results yet very consistent patterns, might contribute to raters generally being more careful and conservative in their decision-making and maybe trying to imagine how someone else might rate a performance.

7.4.5 Rating Behaviour Metrics – Discussion

Deliberation time. There were no statistically significant correlations between total deliberation time and any of the rating quality measures or independent variables. It is noteworthy, that deliberation time correlated positively with the *rational* DMS ($r_s = .37$, ns) and negatively with the *spontaneous* DMS ($r_s = -.36$, ns). These findings are in line with expectations based on DMS theory. Rational decision makers might tend to take in more information and take longer weighing the numerous cues before coming to a decision (Horstmann et al., 2009), while spontaneous decision makers would tend towards faster processing of the information gathered in the performances (Scott & Bruce, 1995).

The decision to include rater behaviour metrics into the current study goes back to Davis' study (2008, 2015). He explored the use of exemplars and average time spent on a performance and found that the more capable raters took longer for their rating decisions (2015, p. 163). The results in the current study neither confirm nor refute his findings and must be considered inconclusive. Davis' results might have been clearer cut as the raters in his study were given more liberties and could decide for

themselves whether they wanted to listen to a performance twice or consult exemplars before entering a decision.

Time to first decision. The analysis also revealed no significant associations between time to first decision and rating quality. There was one moderate negative correlation with accuracy in the criterion TA ($r_s = -.44$), indicating that earlier first decisions coincided with higher accuracy. This finding might seemingly contradict other studies claiming that delaying judgement was conducive to rater quality (Davis, 2008) and even a feature of rater expertise (Wolfe, 2006). However, rater reliability studies such as Wolfe's observed raters during the rating of written texts. When rating spoken language and longer performances, taking stock of one's representation of a performance and taking down a decision at an earlier point might have supported rating accuracy, possibly because this might free up additional resources to consider other criteria.

Revisions. The total number of revisions coincided significantly with accuracy on TA ($r_s = .56$). One possible explanation could be that those raters who particularly made use of the opportunity to revise engaged more deeply with the rating process and were more conscientious and thorough. Being willing to revise decisions might also have been a concomitant of flexible thinking and being able to revise one's impression of a performance as it unfolds which could also contribute to better rating quality. Even though not statistically significant, it is interesting to note that the number of revisions also correlated positively with the *dependent* ($r_s = .29$) and *avoidant* ($r_s = .25$) DMS, which could indicate that a reluctance to rate or lower confidence contributed to a higher number of changes in rating decisions.

To sum up the previous discussion, almost all the associations explored in Study 2 were very weak or negligible, and non-significant. The decision to explore the various cognitive and psychological attributes was based on various cognitive validation studies as well as Baker's (2012) exploratory study of DMS in the context of writing and Davis' (2008, 2015) study of rater training and expertise in the context of speaking. The results in the current study invite the conclusion that the effects of the variable factors as operationalized and investigated were either negligible or very weak or could not be detected adequately with a sample size of 39 participants. Another angle to viewing the results is that the contribution of various attributes to the rating process is complex and potentially mediated by observed or unobserved factors. This leads us to discussing Study 3, which aimed to investigate how the various observations collected and discussed in Study 1 and 2 could be collated for four raters and provide a more nuanced view as to how various features of the rater, the scale and the context bear their influence on rating quality.

7.5 Study 3: Rater Case Studies

7.5.1 Summary

This chapter focussed on four individual raters and investigated how particular attributes and behaviours might shape the outcomes of rating processes. Four raters were selected primarily on basis of their rating accuracy but also their decision-making profiles and the quality of their handwritten notes. The group-level metrics presented for Study 1 and Study 2 were broken down to individual-level metrics and supplemented with further observations and excerpts from the handwritten

notes, responses to the open survey items and the exit question. The contrast between the raters highlighted the individual differences in how raters handled the task of rating spoken performances in real time and, furthermore, revealed the complex interactions of these attributes.

7.5.2 Discussion

The case study approach revealed differences and communalities between the four raters which can also be related to findings in other rater cognition studies. Stormi and Lexi, the two highly accurate raters, were both relatively fast in submitting their final rating decisions in terms of deliberation time. This finding is not congruent with Davis (2008, 2015) who found that raters who took longer were more likely to be accurate. However, he could not explain exactly why this was the case and reasoned that it might have to do with raters taking extra time to listen to certain performances twice or return to the exemplars. In the current study, raters were clearly instructed to watch the performance only once and each rating session was invigilated. While faster to submit their decisions, Stormi and Lexi differed slightly as to when they entered their first decision. Stormi was a more flexible decision-maker as measured in the cognitive tests and entered her decisions early, but also changed them frequently as the performance went along. Lexi, on the other hand, slightly held back with entering her first decision, but also changed slightly fewer of her decisions. The two other raters, Betty and Mary both took considerably longer to enter their first decisions and to deliberate their ratings overall. Once they had entered a decision, they would hardly ever change it.

The handwritten notes and differences in the features attended to, as well as the level of detail provided, added another layer to the raters' profiles. Stormi's and

Lexi's notes were systematic, but at times short, and included verbatim quotes from the performances as well as self-generated statements and global observations that reflected their judgements about the candidate's ability in their own words. Mary's and Betty's notes, on the other hand appeared dominated by detail orientation. These findings are in line with Zhang (2016) and Wolfe (2006) in the context of assessing writing who both observed that more accurate raters are more robust in their understanding of the construct and less drawn towards local features like mistakes or a particular word choice. Zhang (2016) also found that the accurate raters in her sample would summarize salient features more effectively during reading the performances, while the less accurate raters would not summarize until having read the entire performance. She argued that they did so in order to be able to integrate new information as they read the performances. Lexi and Stormi, who both displayed engagement in experiential processing and intuitive decision-making, also appear to have integrated noteworthy features as they listened to the performances. They were also faster than their peers, Betty and Mary, who both captured local detail or no detail in their notes and deliberated longer after the performance had already finished. Stormi and Lexi might partly have been swifter in their judgement because they had already formed a fuller representation of speaker ability during the performance and were possibly also more confident about their decisions. Mary and Betty, might have spent more time reflecting on the performances, weighing the various observations and notes in relation to the descriptors once the performance was over. At least in Betty's and Mary's case, longer deliberation after the performance did not contribute to more effective rating. Interestingly, Wolfe's (2006) finding that accurate raters would adhere closer to the wording of the scale was not confirmed by Lexi's and Stormi's case. This might be

due to the sample in Wolfe's study, who might have been quite experienced given the context of his research, or due to a method effect as Wolfe's analysis was based on think-aloud protocols rather than notes.

Even though both raters were inaccurate and overly severe, Mary's and Betty's understanding of the rating task was quite different from one another, and their inaccuracy might be due to different reasons. They both remarked that paying attention particularly when rating accuracy was challenging for them. However, while Mary struggled to process all the information available, Betty's processing tended to be superficial. Both cases illustrate the discrepancies that can be observed between what Lumley (2005) referred to as "accessible and inaccessible thoughts", or what JDM researchers would call conscious and unconscious processing. As raters engage in the complex rating process and navigate the numerous demands on their attention, their focus consciously or subconsciously shifts towards specific features of the performances and interpreting these features in light of the scale. Particularly in Betty's case automated or at least not consciously perceived processes shape rating decisions more than the ones that are deliberately documented in her notes. Unconscious attention to certain features has been documented in various rater cognition studies (e.g., Lumley, 2005; McNamara, 1996; Zhang, 2016). However, it is important to note that unconscious processing as such does not constitute a threat to rater quality as it has also been observed with highly functional expert raters (Lumley, 2005).

Finally, all raters except Mary, who wrote a lot to begin with, tended to produce more elaborate notes as the experiment progressed and while Mary and Betty might have paid more attention to topical development right from the beginning, Stormi

and Lexi included TA in their considerations as they became more experienced with the rating procedures. Davis (2015) made a similar observation in his study and even found that comments about topic development became more frequent than comments about language features in the later rating sessions. This trend coincided with lower accuracy, leading him to speculate that the increased focus on topic came “at the expense of attention to language issues” (2015, p. 169).

All four raters mentioned at least once that they struggled with the stress of cognitive load as well as dividing and directing their attention. Mary and Betty consistently appeared suspicious of their capacity to detect errors or evaluate what can still be accepted for the criteria RSL and ASL. Stormi expressed such concerns only with relation to TA and Lexi linked feeling overwhelmed with the sheer number of criteria. Having to confirm the correctness of the language produced by the speakers demands activating declarative memory which can impede the processing of the performances considerably and could explain some of Mary’s and Betty’s problems. We can conclude that the challenges of real-time processing were a problem for all raters, but they compensated for the pressures by either resorting to more experiential processing mediated by more global observations and metacognition, or, in the case of Mary and Betty, by exhaustive note-taking or shallow processing. In addition to the general cognitive demands, it is also possible that Mary and Betty have weaker language skills than Lexi and Stormi, and still perceive themselves as learners with a keen focus or ear for mistakes.

One final theme that became visible through the case studies could be the potential role of flexibility. Both, Mary and Betty appeared rigid in the features they noticed in the performances and their notetaking. Stormi and Lexi on the other hand

appeared less structured and displayed a more flexible approach. Despite receiving the same training, Stormi and Lexi acted more similar to expert raters than Mary and Betty: they appeared somewhat authorial in their comments; they were faster in their decision-making; their notes and comments integrated various features of the performance into a more global judgement; and they were highly accurate and consistent (e.g., Lumley, 2005; Zhang, 2016; Wolfe, 2006). It is therefore likely that there may be differences in how these raters structured their knowledge around the rating task and their concept of what a rater is supposed to do.

One way of looking at these behaviours might be through the lens of Cognitive Flexibility Theory (CFT; Spiro et al., 1988). CFT seeks to provide a theory of learning and instruction that fosters individuals' ability to develop a deep understanding of complex subject matters, use and apply knowledge flexibly in dynamic real-world settings, and challenge underlying views and rigid thinking patterns (Spiro et al., 2003). CFT may be applicable to the challenges of training successful rater cognition in that it aims to improve current educational models and promote complex learning and accelerated expertise acquisition. Stormi and Lexi both appear to display features of cognitive flexibility as they avoid “atomistic decomposition of complexly interacting information” and selectively “use . . . knowledge to *adaptively fit* the needs of understanding decision making in a particular situation” (Spiro et al., 1990, p. 5, emphasis in the original).

The intriguing question at this point is why such differences between the two rater pairs emerged despite quite rigid recruitment and standardized training. For one, Stormi and Lexi might have been flexible thinkers to begin with and this construct might have been captured by the *intuitive* DMS score, explaining the effect of the

intuitive DMS on rating quality. However, the four case studies appear to reveal the complexity of the factors involved in rater cognition and how they might interact to produce certain effects (e.g., Mary's need for detail and strained capacity; Betty's confidence and perceived ease of rating paired with a shallow processing of the scale). This underlying complexity explains the somewhat unclear results of studies such as Baker's (2012), or in Study 2 of this dissertation and highlights that rater cognition research needs to consider the interaction of rater attributes.

7.6 Theoretical implications

The overarching research aim of this dissertation was to investigate the nature of rating spoken language.

One of the most widely used basic models of the rating process was proposed by McNamara (1996) and incorporated several key components of the rating process: candidate(s), interlocutor(s), task, performance, rating scale and rater(s) (see also Fulcher's model of speaking assessments in Figure 2.1). The current study focussed on expanding common notions of rater attributes by including measures of cognition, information processing preferences and decision making as well as exploring the interaction of raters with the rating scales and performances. The following wider discussion will first focus on raters and rating scales as central components of performance assessment (Lumley, 2005) before discussing the results as they relate to the rating process.

7.6.1 Raters

This study contributed to the existing body of rater cognition research in that it focused on the rating of spoken language by novice raters. It also investigated the potential of broadening the scope of rater attributes that can be considered as influential on rating accuracy and behaviour by looking at cognitive and psychological aspects. Research into rater cognition in its narrower sense has often explored the features that raters do or do not notice about language performances (e.g., Barkaoui, 2010b; Brown, 2003; Cai, 2015; Cumming et al., 2002; Davis, 2015; Eckes, 2012; May, 2006; Zhang, 2016). In a more encompassing sense, rater cognition research attempts to link individual attributes with rating outcomes and behaviours (Bejar, 2012; Eckes, 2012; Wolfe & McVay, 2012). As discussed in the literature review chapter of this dissertation, raters' first language, second language competence, expertise and training are among the commonly investigated attributes that may lead to rater effects. The current study is among the first to research the effects of decision-making style (with Baker, 2012) and explore the potential role of cognitive attributes and cognitive preferences.

In this study, neither cognitive attributes linked to working memory nor cognitive processing preferences could be identified as consistent or significant predictors of rating behaviour or rating quality. This suggests that these factors may not have a strong effect in rating processes. In addition, analyses of individual cases indicate that other variables (e.g., cognitive flexibility) and metacognitive strategies that raters employ consciously or subconsciously might be effective mediators of limited cognitive capacities. In order to better understand the cognitive processes involved in rater cognition it therefore seems essential to investigate the nature of

these regulatory processes, how they interact with other components of the rating model (i.e., rater attributes, rating scale and context), and how rating procedures as well as rater training may support raters to more effectively activate them.

Decision-making styles might be linked more clearly to rater accuracy, severity, and rating behaviours. Thus, this study expanded the application of this branch of JDM research as suggested by Baker (2012) and confirmed that there is potential in considering decision-making style as one aspect of individual differences in rater cognition research. As a measure of applied decision-making the GDMSI appears to tap into characteristics that shape the process as well as the outcome of rating speaking performance. While it is understood that the rating process is complex and can never be fully captured by listing all features that candidate(s), rater(s) or even interlocutor(s) might bring to the situation (Lumley, 2005), this study contributes to the current body of knowledge that psychological attributes such as decision-making styles might merit inclusion or consideration in future research to help model the rating process and explain rating decisions.

7.6.2 Rating Scale

Barkaoui (2010b) found that it was the type of scale (holistic or analytic) rather than the level of experience that appeared to shape rating outcomes. The development of the Austrian rating scale, which was used in this study, followed common standards of good practice by involving experienced teachers and testers, comparing the scale against performances (Holzknecht et al., 2018), and MFRM-based analyses of rating patterns from several rating workshops (C. Spöttl, personal communication). As Fulcher (2003) argued, investigating the scale use of novice raters may be another possibility to investigate scale functioning. The results from

this study show that even novices were able to achieve acceptable levels of intra-rater reliability (as in MFRM fit) as well as some degree of inter-rater agreement and consistency. However, many participants reported struggling with the scale and the criterion TA consistently emerged as somewhat problematic.

Leaving aside contextual considerations of the scale development, this study nonetheless raises the question of whether the current analytic rating scale optimally serves its prime purpose of helping teachers to consistently assess their students in the final speaking examination. Analytic rating scales have been found to help raters focus on various aspects of the construct, support intra-rater consistency and be particularly helpful for novice raters (e.g., Barkaoui, 2010a, 2010b). On the other hand, fully engaging with every criterion and descriptor in the rating scale takes effort and might be particularly challenging in the context of speaking. The case studies showed that raters aiming for transparent linking with the scale descriptors (Betty) or meticulous notes (Mary) were less accurate than raters that appeared to take a more global approach. Following the recommendation of the CEFR (Council of Europe, 2001, p. 38) and in line with many analytic rating scales in use (IELTS, Cambridge English Examination Suite), the Austrian rating scale included four criteria. However, the scales include many described levels and, as some raters noted in their comments, there is considerable detail with many descriptors in some of the criteria. As Lumley (2005) observed in the context of writing, even highly proficient raters never fully internalize the rating scale but instead revisit the wording of the descriptors as they consider and justify their rating decisions. In the case of the Austrian scale, revisiting many descriptors might contribute to increased cognitive load and decreased decision quality particularly for raters with little experience. Whether a similar degree of scoring validity could be obtained with a

reduced number of descriptors, more global or combined descriptors, or fewer criteria would need to be investigated. In the end, this conundrum of providing fuller description of candidate performance features compared to usability concerns goes back to defining clear purposes and aims for rating scales (Alderson, 1991; Fulcher, 2003). After all, the Austrian scale for speaking was deliberately designed to be somewhat similar to the writing scale and communicate the test construct clearly to stakeholders that were not yet very familiar with rating scales. It is therefore not surprising that it does not fully serve any of these purposes and constitutes a compromise that teachers will make work in one way or the other during the examinations and in the classroom. This, too, would merit a closer investigation.

The finding that more successful raters may be applying a more global or configurational approach raises the question of how analytic scales might be used by different kinds of raters. Despite referring to analytic scales, successful raters may not use them in a fully rational, analytic, step-by-step fashion, but rather as means of estimating the quality of separate features within an overall global impression. This is in line with the psychometric assumptions at the basis of MFRM in that multiple cues are combined to locate the position of an individual along the slope of a latent ability variable. The separate criteria and implication to fully compartmentalize each criterion to avoid halo effects, however, might lead to raters applying different approaches based on the same rating scale in that some may shift flexibly between bottom-up and top-down processing of the performance in light of the descriptors while others could be relying more on building a score bottom up from many separate observations. A flexible or balanced approach might be more successful, particularly when it comes to applying knowledge to many different

performances as is hypothesized by Cognitive Flexibility Theory (Spiro et al., 1988). This would need to be investigated more rigorously by looking at how exactly raters might process rating scales during the rating of speaking, but also contrasting different forms of rater training and whether they might impact rater behaviour.

In line with many previous studies about rater cognition, this study showed that not all features included in the test construct are considered with equal scrutiny and severity (e.g., Cai, 2015; Eckes, 2015; McNamara, 1996). Aspects of language use were rated more severely than features such as fluency or content. This raises concern about the inferences that can be made on basis of the scores. If analytic rating scales are implemented with the intention to provide more detailed feedback to learners, the question remains as to how reliably raters can identify jagged learner profiles in the first place. Differences in bands between the criteria might as much be the function of severity patterns than actual differences in the ability of the candidates.

Finally, the criterion TA which was designed to assess fulfilment of language functions and quality of content, repeatedly proved to be problematic from a psychometric perspective. Including this criterion and scaling it for the purposes of the assessment was an important pedagogic signal to teachers and students alike (Holzknecht et al., 2018). However, having integrated the criterion into the scale introduced a different set of problems as raters found it difficult to evaluate the quality of the content while listening to other features. As Fulcher (2003) emphasized, it is more challenging for raters to relate the content quality of individual performances to general scales that apply to a broad set of tasks. Given

the heavy emphasis that was put on content and subject knowledge in the previous form of the *Matura*, it is safe to assume that teachers as well as educational authorities would not have taken on board a rating scale that does not assess the content of a candidates' performance.

7.6.3 Rating Process

For this study, time stamps were recorded with a view to capturing some of the processes taking place during rating. The analysis of time stamp data was inconclusive. This indicated a high degree of intra-individual variation, as was also observed by Davis (2008), in how raters approached the rating of the performances. Furthermore, the somewhat mixed results underline the fuzziness and complexity of the aspects of rating quality that were scaled in the MFRM analyses and operationalized as dependent variables in this study.

One hypothesis from the outset was that a propensity to decide rationally, as operationalized in the decision-making style questionnaire (GDMSI), and preference for rational processing, as operationalized in the rational-experiential inventory (REI-40), could be valuable attributes in raters when applying analytic rating scales to speaking performances. Instead, the results of this study seem to indicate that a preference of rational decision making is unrelated to accurate rating and that a more intuitive and avoidant rating style is associated more distinctly with accurate rating. If intuitive decision-making indeed was an advantage over rational decision making when rating spoken language, this potentially also underlines how the real-time nature of assessment shapes the actual decision process. Rating processes are shaped by shifting and dividing attention between the scale and the performance (listening and reading), between different criteria (e.g., content quality

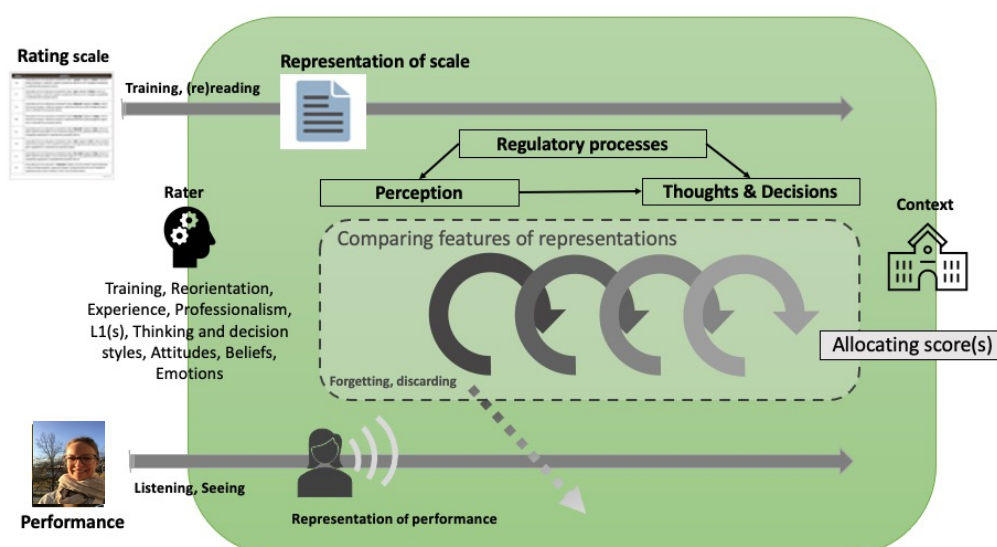
and language quality), and different components of memory (language comprehension and explicit knowledge about language rules and usage, or knowledge of the world). Under time pressure and in the face of complex or non-decomposable tasks (i.e., tasks that do not comprise a distinct set of operations and where analytical thinking is of limited benefit), successful decision-making will rely more heavily on intuition and some form of domain expertise (Dane et al., 2012). While none of the raters in this study could be argued to be domain experts in assessing speaking, Stormi and Lexi might have had a better grasp of the typical features of B2-level English which, paired with a degree of flexibility and intuition, translated into more accurate ratings.

One contribution of the current study is that it conceptualizes the rating task more clearly as a complex cognitive activity that is in itself the performance of a skill shaped by the particular attributes of the raters and brought into action within a certain context. McNamara (1996) suggested about rater-scale interactions modelled via MFRM that they are “like a ‘test’ of the raters (and the scale) in the way that the subject-instrument interaction is a test of the subjects (and of the instrument)” (p. 121). Despite being novices and relatively unfamiliar to both, rating and teaching, the participants in the study succeeded to some extent in using the scale consistently and accurately. Furthermore, conceiving of rating speaking as a non-decomposable task (Dane et al. 2012 also list judging artwork, the taste of food or the difficulty of basketball shots), or as applying knowledge in ill-structured, messy domains (see Spiro et al., 1988) offers a new angle of investigation into rater cognition. This might be particularly beneficial to develop and explore alternative modes of rater training which aim at building a common understanding of the construct via the cognitive flexibility framework.

Previous research has found various models to conceptualize the rating of language performance for the purposes of assessment (e.g., Bejar, 2012; Fulcher, 2003; Lumley, 2005; McNamara, 1997; Weir, 2005; Wolfe 1997). Some emphasize the components, or facets, involved in rating (e.g., McNamara, 1997, Weir, 2005) and provide guidance that is particularly valuable for implementing and validating assessments. Such models can also be extended by including the psychometric measures that may describe and explain some of the relationships between these components (see Eckes, 2015, p. 49). The psychometric framework of invariant measurement (Engelhard & Wind, 2018) integrates a theory of judgement (in this case Brunswick's lens model) and measurement theory. While useful in many contexts and for many purposes, these perspectives do not attempt to model the actual rating process, which is, as Eckes quite rightly argues "far more complex and dynamic" (2005, p. 48) than could be captured by a single diagram. In many professional assessment contexts, there appears to be less of a need to understand the interplay between rater, criteria and performance; raters can be selected based on merit, trained, and retrained, offset via statistical procedures, and even discharged. Figure 7.1 suggest a somewhat expanded component model of rater cognition that draws on McNamara's (1996), Bejar's (2012) and Wolfe's (1997) descriptive models. It is by no means all-encompassing, nor does it fully represent the cognitive structures involved. The main goal is to emphasize more sufficiently than other models the pressures of real-time processing that appear more relevant when assessing speaking as compared to assessing writing. The arrows are meant to symbolize how time and memory as a factor might impact the forming of representations as the speaking performance progresses. It also highlights the distinction of subconscious perception, conscious thoughts and decisions as well as

regulatory processes (i.e., directing attention, allocating resources, metacognition) as conceptualized in human information processing models.

Figure 7.1 Expanded model of rater cognition in rating speaking



In the long run, a more rigorous model of the processes involved in the rating of spoken language performance is needed (Purpura, 2013) and a lack thereof continues to hamper efforts to make the training and improvement of rater performance more effective. Working toward a cognitive model of rating would support the field in moving beyond its current understanding as such models might provide a basis for investigating and testing relationships between components and identify sources of error. Conceptualizing the rating process as a complex task that activates various cognitive components – as is done in other areas of applied psychology (e.g., management, aviation, medicine) – also has the potential to inform the development of rating materials (scales, supplements, exemplars, guidelines), assessment procedures, and rater training. Such an approach could help meet the needs of users of language testing research in educational contexts where raters cannot be selected or released based on their performance.

There are examples of models with real-world applications that encompass memory, but establish its relationships with perception, attention and decision making more clearly, while also including system factors (which could be rating context, rating materials or scales, etc.), and individual factors (rater attributes like L1, L2, training, etc.) (Martins, 2016, in the context of aviation and White et al., 2017, modelled decision-making and skill developed in laparoscopy surgery). In addition to the component of memory such a model would allow more succinct theory building about how certain rater attributes (e.g., social stereotypes, cp. Huhta et al., 2019), cognitive states (e.g., stress, emotion), or rating contexts might shape the processing of the performances and criteria.

While there has been sustained interest in establishing the effect of individual differences on the level of the test taker, the conceptualization of the raters and their attributes remains quite shallow. This becomes apparent when looking at models of speaking test performance such as, for example, the expanded model by Fulcher (Figure 2.1, based on Skehan, 2001), which considers the real-time processing capacities and individual variables like personality for the candidate but more vaguely refers to rater characteristics and training. This view might work in a setting where raters are selected based on completing extensive training and demonstrating considerable merit. However, language teachers cannot and arguably should not be selected based on their ability to rate written or spoken performance, but on their ability to teach language which does include assessment but also an abundance of other skills. If accomplishments in language testing are to be relevant to educational practice, they also need to offer insights and tools that are applicable to pedagogic contexts. For the rating of spoken language this would mean investigating the black box that is rater cognition more extensively with a view to creating assessment

procedures and assessment tools, as well as being able to diagnose problematic rating behaviours, successfully recommend rating strategies and even tailor rater training more specifically to individual needs.

7.7 Implications for a Validity Argument

This thesis looked at novice rater cognition and the potential impact of previously under-researched rater attributes. It was not the purpose of this study to critique or validate the rating scale, rater training procedures or decision processes during the Austrian Matura examinations. In addition, the participants were novice raters with no experience who were provided a standard two-hour training session and the speaking performances were not collected under live test conditions. Nonetheless, following Knoch and Chapelle's (2017) framework, the data from this experiment may be used to lend backing to several assumptions underlying the evaluation inference underpinning rater-mediated assessment (see also Table 2.1 on p. 39 for excerpt).

For the testing procedure to produce “scores with the intended characteristics” (Knoch & Chapelle, 2017, p. 7) and support the evaluation inference, the rating scale needs to be found to function as intended (Warrant A) and the raters need to be found to rate performances reliably (Warrant B). Regarding Warrant A, the data from the MFRM severity and consistency analysis (Section 4.3) suggests that the scale can support raters to distinguish between speakers at different ability levels (Assumption 3 in Knoch & Chapelle's 2017 framework). Despite the limitations of the recorded performances (i.e., small number of performances, relatively narrow range of ability levels included), there were sufficient observations for each step of

the scale above the pass level (Assumption 2). Judging from the written comments regarding the rating scale (see Section 4.6 Perception Data), there were also few indications for overlap between the scale criteria RSL, ASL and FLIN and they are likely to be assessing separate constructs (Assumption 1). However, there were numerous instances in the data that pointed towards issues with the TA criterion, suggesting that this sub-scale may be functioning differently and that raters struggle with rating this criterion. It might also be useful to carry out a factor analysis as suggested by Knoch and Chapelle (2017) to bolster this observation with more empirical evidence.

As far as Warrant B is concerned (i.e., that raters rate reliably), the findings from this study may provide evidence for those assumptions of the framework that strictly pertain to the rating process (for Assumptions 4, 5, and 6). There is some evidence that the novice raters in this study were able to perceive differences across score levels (Assumption 4), but their comments also indicated issues with differentiating between certain ability levels for certain descriptors. This may also have to do with their lack of experience with B2-level language features and might be less pertinent for experienced teachers. Empirical evidence from more experienced raters and on a broader sample would be necessary to see if teachers are confident in allocating scores across all ability levels based on this scale. The results from the MFRM analyses also indicate that the majority of raters were able to apply the scale consistently (Assumption 5). However, around 20 percent of the raters produced mean-square statistics above the recommended threshold levels which indicate issues with rating consistency (see also 4.3.3 Rater Facet Measurement Results, p. 154). The self-report comments regarding the scales seem to suggest that the novice raters were able to gain more confidence in their decisions

as the experiment progressed (Assumption 6, see also Sections 4.6.1 to 4.6.3). Thus, scale users may be able to overcome lack of training with the scale after a certain number of performances.

The findings from this study lend some evidence to several assumptions of the evaluation inference. As will be discussed in the following section, the findings also highlight further steps that could be taken to address some of the remaining assumptions of the evaluation inference (7, 8, 9 and 10, Knoch & Chapelle, 2017, p. 7) that may be met by providing additional and more systematic training opportunities, more detailed support materials and guidance to help create more reliable rating procedures.

7.8 Practical Implications

This study used novice raters with minimal training and outside of the typical context of rating. The novice raters in this study showed varying success in applying the rating scale and disagreed to some extent in their ratings. The claim that each rater is interchangeable which would be the basis of a validity argument cannot be upheld in face of these findings. However, the examination in itself is already embedded in a context and network of legal requirements that currently make implementing a more valid procedure extremely challenging and also unlikely. Nonetheless, it is worth examining which measures could contribute to making the rating of the performances within the given context more reliable.

Given the limited training Austrian teachers receive in the use of the scale and rating in general, the results of classroom assessment as well as examinations are likely flawed. In-service training is voluntary and provided by freelancing teachers which

have usually attended a training course on the rating scale. In addition, speaking assessments in the context of the *Matura* are carried out under conditions that are prone to foster inaccurate or inconsistent rating. The examination is public in that other students, teachers and even parents may attend. In the real exam situation, teachers are required to settle the grade right after the examination and grades must be defended on the spot, sometimes even to exam panel members who are not language teachers. There is no tradition of recording or externally marking the examination, yet both steps could considerably reduce the cognitive load raters experience during the examination and allow raters to consult and confirm their initial rating of the student at a later point. A possibility to reduce error or disagreement due to divided attention is to clarify the role of the second teacher who is usually tasked with using the holistic rating scale. If they were to clearly focus on task achievement, which according to statements from the participants of this study seemed to interfere with the rating of the other criteria the most, this could help reduce errors due to task complexity. However, a procedure for settling disagreement between the raters might have to be put in place.

A survey of Cambridge assessors by Gilbert and Staub (2014) showed that regular examining, face-to-face meetings, and the availability of a range of commented exemplars were among the most useful measures to provide raters with more confidence in their rating. The *Matura* takes place once a year, but it is in the teachers' hands to seek practice and regularly try to apply the scale. After all, deliberate practice is a key component of developing expertise (Ericsson et al., 1993) and this could be included and foregrounded in the materials which accompany the rating scales. Several speaking and writing performances have been published with comments and justifications to illustrate pass performances

(Eberharter & Spöttl, 2018). However, they are limited to pass performances and relatively clear-cut cases. To build their skills as raters, teachers would need to rate and engage with a broader range of performances via an online platform. The approach chosen for creating the benchmarks in this study could be a potential blueprint for a more streamlined procedure of benchmarking and could be extended to systematically collect verbalized observations about the speakers' performances to supplement the fair score. This could help create a larger repository of benchmarked performances to be used by Austria's teachers.

In addition to releasing more annotations to the rating scales and more performances to practice with, efforts could be made to help establish local communities of practice to provide teachers with the support they need to develop their rating skills and create a space where questions and concerns about assessment can be discussed. The materials outlined above are among the basic requirements to achieve this goal, but teachers would need first-hand experience of inhouse benchmarking sessions to see the potential such discussions have for professional development.

Finally, in light of the findings of this study, the question remains whether the typical templates of rater training and instructions given to the raters need to be revisited. For instance, the typical protocol of step-by-step, or criterion-by-criterion familiarization with the rating scale followed by the rating of an illustrative performance, comparison with benchmark and discussion could be explored and re-evaluated. This procedure might suggest that compartmentalized analysis fosters the best rating results; however, it might not optimally prepare raters for the rating of spoken language. Recent technological advances – including increasing band-

width capacities in many countries – would allow to create more immersive, non-linear and multidimensional learning experiences that might be more conducive to fostering rich representations of the construct and the necessary flexibility to apply this understanding to a broad range of performances (Spiro et al., 1988; 2003). Furthermore, as Raczynski et al. (2015) suggest, self-paced training may produce similar results than resource-intensive collaborative training sessions.

7.9 Research Method

The main feature of this study that might be useful for other studies in rater cognition research is the use of response time data via time stamps. Despite the mixed results, the time stamp data provided insights into how raters engaged with the performances. First, the time stamps increased the trustworthiness of the rating scores in that they helped to confirm that raters were indeed watching and rating the performances with due diligence. This was very valuable given that the raters were working independently and without close observation at the time of data collection. Analogous to test taker response data, which is increasingly explored for its potential to shed light on item difficulty, test-taking processes, and validation (e.g., Ercikan & Pellegrino, 2017; Zumbo & Hubley, 2017), response data collected via online platforms could also be considered an interesting avenue for researching experimental as well as live test administration sessions. As is the case with test taker response data, numerous studies would be required to be able to interpret these metrics in relation to rating processes (e.g., Li et al., 2017). Even in an experimental setup there is ambiguity as to what aspects of rater cognition are captured by these measures. However, while process data such as time stamps could certainly be useful in detecting problematic or interesting rating behaviour to investigate further,

it is contingent on what raters do while rating (note-taking, deliberation, viewing exemplars, etc.).

Resorting to the RAM (rater accuracy model) based on accuracy scores rather than the raw scores in the MFRM analysis was useful for several reasons. First, rather than having to agglomerate various measures of rater accuracy or deciding on a particular measure based on CTT, the MFRM-scaled measures provide a theoretically sound compromise. Second, the MFRM based accuracy measures are useful conceptually as they provide a deeper understanding of scale functioning through criterion and bias reports. It is frequently observed that severity is less impacted by training and expertise than accuracy, which would be an argument to resort to the RAM model more frequently to evaluate the effectiveness of rater training. A downside to MFRM accuracy measures, however, is that the transformation of raw scores to accuracy scores makes them more muted than severity measures which makes detecting effects more difficult, as was attempted in Study 2.

8 Conclusion

This study set out on the premise that rater variability and the rating process are still undertheorized in the context of assessing speaking. By delimiting the sample to novices and a quite homogeneous group in terms of rating and teaching experience, it examined rater cognition by focusing on: 1) features of rating quality in novice raters, 2) the potential role of attributes such as cognitive capacity, preferred processing mode and decision-making style, and 3) how individual raters with different accuracy profiles differed in terms of their behaviour and experience during the rating task. The previous chapters (4-6) presented the results for each research focus. Chapter 7 summarized and discussed the findings in light of the research literature and considered the broader theoretical, practical and methodological implications for rater cognition research. This concluding chapter will first present some reflections on the limitations of the study and then close with suggestions for future research endeavours.

8.1 Limitations

This study has several limitations which need to be considered when interpreting the findings presented in the earlier chapters. First and foremost, to be able to isolate the effects of certain components of the rating process, it was necessary to reduce the number of variables that would co-occur in more natural assessment settings. While necessary for the purposes of observation, this reduction of variables (e.g., tasks, forms of interaction, interlocutors, settings) naturally limits the inferences that can be made from this experiment to actual real-world rater behaviours. As Fulcher (2003) explains:

our ability to speak does ‘change’ depending not only upon whom we are speaking to, where and about what, but under what conditions. With every change in each variable of the context in its broadest sense, the scores may also change. (p. 138)

For instance, this study focused on just one task and one task type which was the monologue. Assessing numerous test takers on a reduced number of prompts may be typical for centrally delivered examinations but is not the actual model for the Austrian *Matura*. In the real speaking examination, the final score is determined by combining the scores from two teachers – one using the analytical and the other using the holistic scale – and two task formats – the individual task and the paired activity. Therefore, at least one rater will have to assess a speaker in both modes and the load of assessing two speakers may be quite different to assessing just one. In addition, the performances were recorded to be able to select and combine them based on fair scores and present them in a randomized design. This will likely have had an impact on the speakers, who did not perform under real test conditions, as well as the raters, who might rate differently when in the same room with the candidates.

Another limitation related to the experimental nature of this research may be the sampling of speakers and raters. As far as the speakers were concerned, not all were preparing to take the speaking examination which meant that the speakers differed in their preparedness for the task format. While the degree of readiness for an examination may vary in actual examinations as well, it certainly impacted the quality of performances in terms of length and at times required additional prompting from the interlocutor which in turn might have impacted ratings of task achievement. A related problematic factor of the sample performances was the quite narrow range they represented in light of the rating scale as only very few reference

scores were beneath the pass band 6. This might have impacted raters in that some of them may have expected a greater range and consciously or subconsciously amplified differences between the performances to try spread them more across the scale bands.

As far as the raters were concerned, several steps were taken to reduce the effect of teaching and rating experience. However, even though all raters could be considered advanced in their language skills and the number of semesters was not found to be a factor, the case studies indicated that language proficiency might still have played a role. It is certainly the case, that language proficiency will fluctuate in the broader population of English teachers. In the absence of external measures to truly estimate the raters' language proficiency the potential effect of this variable cannot be determined with certainty, but language proficiency could be affecting rating quality directly, if raters must activate explicit knowledge structures to assess language accuracy, or indirectly, if raters' still perceive themselves to make many mistakes and are attuned to detecting errors.

Finally, measuring rating accuracy in this study depended on comparing scores with reference scores from expert raters. Even though great care was taken to select experts and moderate their ratings via creating MFRM-scaled fair scores, the reference scores might not fully reflect the scores an actual benchmarking panel might create for the performances.

Although the raters of this study attended a rater training session, they may also have taken some of the more pedagogic aspects of rating spoken language under consideration while evaluating the performances and the findings from this study might not predict very well how the raters may rate their own students in a

classroom setting. Assessing speaking is just one of many complex tasks and interpersonal interactions teachers deal with in the context of an English learning classroom. To operate successfully, there are many tasks future teachers need to be able to deal with and assessment within classroom settings is likely impacted by numerous factors that were not considered in this study. Among these factors are the teachers' familiarity with their students and their individual ways of dealing with tasks and speaking. In addition, any assessment is set within a particular context that may vary considerably for each teacher, student, classroom, school and even region.

It should also be noted that the procedures of data collection in themselves might have had an impact on the findings. In this study, the participants rated only a limited number of performances under proctored conditions. While they were allowed and encouraged to take as many breaks as they needed, they might have decided to continue rating despite being tired. Therefore, fatigue might have played a certain role and affected the results. Moreover, the number of performances (i.e., 15 performances of 5 minutes each per rating session) is certainly higher than would be the case on a typical testing day at an Austrian school. On the other hand, in the real-world context teachers must assess over several days, under high time pressure with their peers, the school head and external officials in the room. Teachers may also have to switch between assessing different subjects or between different schools during this time of the year. Even though the experiment in this study was certainly tiring and somewhat strenuous for the participants one could argue that the pressures of assessing in the natural setting might be higher.

The method chosen to collect the qualitative data may also have had an impact on the generalizability of the findings. Previous rater cognition studies have often employed encompassing methods like think-aloud protocols and written decision-making records (see Baker's write-alouds). In many cases, these methods were more suitable or manageable because the project focussed on assessing writing, which allows more easily for concurrent introspection. Davis (2008), for instance, collected a lot of qualitative think-aloud data which in the end could not be analysed extensively within the remit of the study. Therefore, the decision of collecting Likert-type responses and short justifications in combination with the handwritten notes constituted a workable compromise which may, however, lack fine-grained information. Furthermore, many rater cognition studies are based on rather short performances or holistic rating which makes it easier to increase the number of rating decisions that can reasonably be gathered from participants. The scores in such designs might be a more robust representation of rating quality as ratings might be more stable over a larger number of performances. The decision to collect the ratings for 30 speakers was within reason for the participants and robust enough for MFRM analysis (Linacre, 2020, or Eckes, 2015, recommend a minimum of 30 observations per element).

Finally, the sample size of 39 raters was feasible for a PhD project, but not powerful enough to reliably detect weak effects or model interactions. It also limited the range of context dependent and independent aspects that could potentially be explored in their relation to decision-making (Appelt et al., 2011). The decision-making measures and cognitive ability measures included in this study offer a limited glimpse. Other measures that could be revealing of underlying cognitive processes are constructs such as motivation or the role of personality, either

operationalized in terms of an inventory (e.g., the Five Factor Theory of Personality) or targeted through various distinguishable subconstructs (e.g., empathy, impulsiveness, perceived locus of control). Even though decision-making styles are an area of applied psychological study that is seeing growing interest, particularly for its potential to support professional development, the constructs supporting measures such as the GDMSI are fuzzy, undertheorized and require more validation. Consequently, as the discussion of the results of Study 2 has shown, decision-making styles might offer interesting insights and could serve as proxies to explore constructs that are possibly even more difficult to operationalize, but they also need to be interpreted with care.

8.2 Recommended Further Research

Looking forward, further research could address some of the limitations just discussed to confirm some of the findings or narrow down some of the issues raised.

First of all, the stability of the findings from this study could be confirmed with a larger sample. It would be interesting to examine how the questionnaire data behaves in a broader sample of experienced raters and whether similar patterns emerge with measures of rating quality despite extensive rater training and experience. Moreover, it might be worthwhile to collect data on cognitive tasks in a larger sample despite the mixed results in the current study. It appears unlikely that cognitive factors play a large role in the rating process. However, if the sample is reasonably large the data on cognitive tasks could be of theoretical interest in that it could provide insight into which features of cognition are the most influential when it comes to rating speaking. As the development and validation of remote

cognitive testing methods thrives during the current Covid-19 crisis there is reason to believe that such a study could be implemented even in a larger sample of raters.

In the current study, the rating process itself was partly reconstructed by rater behaviour metrics and handwritten notes. This approach revealed insights into how raters may be processing the performances and rating scales, but studies that employ a more fine-grained approach are needed to confirm, for example, whether there are different engagement patterns with the rating scale for different groups of raters defined by rating behaviours or context independent traits. An eye-tracking study combined with stimulated recall could help construct a much fuller picture of what the rating process might look like for different raters. Following the recommendations of Banyard and Svenson (2011) for modelling complex decision-making processes, such an endeavour should set out by formulating general expectations of the process, describing task demands, defining seminal moments during the process, and matching the observations from the eye-tracking with comments from the stimulated recall and expectations. Because raters are confronted with so many decisions when rating with analytic rating scales, it might also be of interest to design studies that partly decompose analytic rating and contrast the rating processes of rating by criterion with rating all criteria. This might shed more light on the costs and gains of rating analytically as well as the raters' mental representation of the performances and rating task in any particular condition. However, as the findings from this study highlight, there are substantial inter-individual differences in how rating is approached that might be determined by numerous factors. While no such research study on its own will fully explain the rating process, it may shed more light on the potential weaknesses of rating procedures and unsuccessful – or successful – strategies raters of speaking employ.

Given that intuitive thinking and flexibility might play a role in rating speaking, future studies could look at rater training, and the potential effects of training procedures and instructions on rater behaviour. This would forgo investigating the rating process from a cognitive perspective and instead focus on how to influence the outcome of rating processes. As discussed previously, rater training appears to follow a typical protocol which introduces raters to the rating scale in a step-by-step fashion. When it comes to rating all criteria, which is the ultimate goal of the training, raters are possibly left to their own devices to adapt successful strategies and cope with the demanding task. One approach to reaching a certain point of expertise is providing many performances. However, rather than emphasizing the duration of practice the emphasis could also be put on the quality and depth of practice. Therefore, gauging the potential benefits of Cognitive Flexibility Theory (Spiro et al., 1988) and looking into developing more immersive, self-paced and reflective forms of rater training and comparing their effect on the development of rating quality could be useful for the language testing community. Again, however, the success of any training method might be moderated by individual preferences and needs, making surveying such individual factors an important component for such studies.

Finally, it could be of interest to investigate how the rating scales are in fact used by Austrian teachers in their pedagogical practice. Theoretically, English teachers could use the B1 and B2 assessment scales throughout the last four years of upper-secondary education. However, I am not aware of any research that examines the uptake or use of these scales in the context of classroom assessment. This could even be extended by looking at the use of the writing scales, which have a parallel structure and must be used by all teachers for the compulsory final written

examination. Such a study could look at levels of training received, attitudes towards the scales and particular criteria, but also how teachers perceive the use of the scales when assessing writing as compared to speaking. Insights from such a study could deliver valuable information for scale revision as well as a clearer baseline for teacher training needs.

8.3 Conclusion

The goal of this study was to examine the nature of the rating process when assessing spoken language. The findings of this study contribute insights into the experience of the rating process from the perspective of novice raters and the potential contribution of cognitive and psychological rater attributes which previously have not been considered extensively within the field. The results suggest that research into rater effects might have to acknowledge the possibility of interaction effects of various rater attributes and investigate the potential of metacognition and flexible thinking. In this study, rating emerges as a complex, dynamic and individualised process; future research needs to embrace this complexity if progress is to be made towards a comprehensive theory of rater cognition.

9 References

- Ackermann, K., & Kennedy, L. (2010) Standardizing rater performance: Empirical support for regulating language proficiency test scoring. *Pearson Research Note*. https://pearsonpte.com/wp-content/uploads/2014/07/RN_StandardizingRaterPerformance_2010.pdf
- Aickin, M. & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American Journal of Public Health*, 86(5), 726–728. <https://doi.org/10.2105/AJPH.86.5.726>
- Akinci, C., & Sadler-Smith, E. (2013). Assessing individual differences in experiential (intuitive) and rational (analytical) cognitive styles. *International Journal of Selection and Assessment*, 21(2), 211–221. <https://doi.org/10.1111/ijsa.12030>
- Alderson, J. C. (2011). Principles and Practice in Language Testing: Compliance or Conflict? Presented at the IATEFL Testing and Assessment Special Interest Group, Innsbruck.
- Alderson, J. C. 1991: Bands and scores. In Alderson, J. C. & North, B., (Eds.), *Language testing in the 1990s* (pp. 71–86). London: Modern English Publications/British Council/Macmillan.
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

- Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1–15.
<https://doi.org/10.1146/annurev.ps.37.020186.000245>
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–573.
- Appelt, K. C., Milch, K. F., Handgraaf, M. J. J., & Weber, E. U. (2011). The Decision Making Individual Differences Inventory and guidelines for the study of individual differences in judgment and decision-making research. *Judgment and Decision Making*, 6(3), 252–262.
- Attali, Y. (2015). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, 1–17.
<https://doi.org/10.1177/0265532215582283>
- Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press.
- Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing*, 17(1), 1–42.
<https://doi.org/10.1177/026553220001700101>
- Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.
- Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12(2), 238–257.
<http://doi.org/10.1177/026553229501200206>

- Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11), 417–423.
[https://doi.org/10.1016/s1364-6613\(00\)01538-2](https://doi.org/10.1016/s1364-6613(00)01538-2)
- Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, 4(10), 829–839.
<https://doi.org/10.1038/nrn1201>
- Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63, 1–29. <https://doi.org/10.1146/annurev-psych-120710-100422>
- Baddeley, A., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, 8, 47–89. [http://dx.doi.org/10.1016/S0079-7421\(08\)60452-1](http://dx.doi.org/10.1016/S0079-7421(08)60452-1)
- Baiocco, R., Laghi, F., & Alessio, M. D. (2009). Decision-making style among adolescents: Relationship with sensation seeking and locus of control. *Journal of Adolescence*, 32(4), 963–976.
<https://doi.org/10.1016/j.adolescence.2008.08.003>
- Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly*, 9(3), 225–248. <https://doi.org/10.1080/15434303.2011.637262>
- Ballard, L. (2017). *The Effects of Primacy on Rater Cognition: An Eye-Tracking Study* (Publication No. 10274418) [Doctoral dissertation, Michigan State University]. ProQuest Dissertations Publishing.
- Banyard, R., & Svenson, O. (2011). Verbal data and decision process analysis. In M. Schulte-Mecklenbeck et al. (Eds.) *A Handbook of process tracing methods for decision research: A critical review and user's guide* (pp. 115–137). New York and Hove: Psychology Press.

- Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, 44(1), 31–57. <https://doi.org/10.5054/tq.2010.214047>
- Barkaoui, K. (2010b). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7(1), 54–74. <https://doi.org/10.1080/15434300903464418>
- Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, 18(3), 279–293. <https://doi.org/10.1080/0969594X.2010.526585>
- Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement Issues and Practice*, 31(3), 2–9. <https://doi.org/10.1111/j.1745-3992.2012.00238.x>
- Berisha, G., Pula, J. S., & Krasniqi, B. (2018). Convergent validity of two decision making style measures. *Journal of Dynamic Decision Making*, 4(1), 1–8. <https://doi.org/10.11588/jddm.2018.1.43102>
- Bernard, H. R., & Ryan, G. (1998). Text analysis: Qualitative and quantitative methods. In H. R. Bernard (Ed.), *Handbook of methods in cultural anthropology* (pp. 595–645). Walnut Creek, CA: AltaMira Press.
- BMBF, Bundesministerium für Bildung und Frauen (2012). *Die kompetenzorientierte Reifeprüfung: Lebende Fremdsprachen – Unterlagen zur Beurteilung*. https://www.bmbwf.gv.at/dam/jcr:f02885f2-e343-4441-a710-36b0aae305a9/reifepruefung_ahs_lflfpub_24029.pdf
- BMBF, Bundesministerium für Bildung und Frauen (2013). *Die kompetenzorientierte Reifeprüfung: Lebende Fremdsprachen – Richtlinien*

und Beispiele für Themenpool und Prüfungsaufgaben.

https://www.bmbwf.gv.at/dam/jcr:0c6fc255-08bf-4495-ba35-b7fa3dcd8769/reifepruefung_ahs_lflfsp.pdf

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model* (Third edition). New York, London: Routledge.

Bönisch, M., Maldet, A., & Zheng, J. (2020). *Standardisierte Reife- und Diplomprüfung: Abschlussjahrgang 2018/19 Sommertermin Tabellenband*. Wien: Statistik Austria.

http://www.statistik.at/wcm/idc/idcplg?IdcService=GET_PDF_FILE&dDocName=122942

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.

<http://dx.doi.org/10.1191/1478088706qp063oa>

Bröder, A. & Newell, B. R. (2008). Challenging some common beliefs: Empirical work within the adaptive toolbox metaphor. *Judgement and Decision Making*, 3(3), 205–214.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12(1), 1–15. <https://doi.org/10.1177/026553229501200101>

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. *IELTS Research Reports, Volume 3*. IELTS Australia. <https://www.ielts.org/research/research-reports/volume-03-report-3>

Brown, A., & Taylor, L. (2006). A worldwide survey of examiners' views and experience of the revised IELTS Speaking Test. *Research Note*, 26, 14–18. <https://www.cambridgeenglish.org/images/23145-research-notes-26.pdf>

- Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tasks. *TOFEL Research Report*. Educational Testing Service. <https://www.ets.org/Media/Research/pdf/RR-05-05.pdf>
- Brown, J. D. (2005). *Testing in language programs* (New edition). New York: McGraw-Hill.
- Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, 92(5), 938–956. <https://doi.org/10.1037/0022-3514.92.5.938>
- Brunfaut T., & Harding L. (2018). Teachers Setting the assessment (literacy) agenda: A case study of a teacher-led national test development project in Luxembourg. In D. Xerri & P. Vella Briffa (Eds.), *Teacher Involvement in High-Stakes Language Testing*. Springer, Cham. https://doi.org/10.1007/978-3-319-77177-9_9
- Brunfaut, T., & Révész, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly*, 49(1), 141–168. <https://doi.org/10.1002/tesq.168>
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1), 116–31. <https://academic.csuohio.edu/kneuendorf/quillin/cacioppo%20petty%20the%20need%20for%20cognition%201982.pdf>
- Cai, H. (2015) Weight-based classification of raters and rater cognition in an EFL speaking test. *Language Assessment Quarterly*, 12(3), 262–282. <https://doi.org/10.1080/15434303.2015.1053134>

- Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, 28(2), 201–219.
<https://doi.org/10.1177/0265532210393704>
- Casanave, C. R. (2015). Case studies. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource*. (pp. 119–136). London, New York: Bloomsbury.
- Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19. <http://doi.org/10.1017/S0267190599190135>
- Chapelle, C. A., Chung, Y. R., Hegelheimer, V., Pendar, N., & Xu, J. (2010a). Towards a computer-delivered test of productive grammatical ability. *Language Testing*, 27(4), 443–469.
<https://doi.org/10.1177/0265532210367633>
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as Foreign Language* (pp. 1–26). New York & Oxon.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2010b). Does an Argument-Based Approach to Validity Make a Difference? *Educational Measurement Issues and Practice*, 29(1), 3–13.
<https://doi.org/10.1111/j.1745-3992.2009.00165.x>
- Colman, A. M. (2015). Product–moment correlation coefficient. In A. M. Colman, *Dictionary of Psychology* (4th ed.) Oxford: Oxford University Press.

- Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and Conducting Mixed Methods Research* (Third edition). Thousand Oaks, CA: Sage.
- Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice*, 31(3), 10–20.
<https://doi.org/10.1111/j.1745-3992.2012.00239.x>
- Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence Measurement, Theory, and Public Policy* (pp. 147–171). Urbana: University of Illinois Press.
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper and Row.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. <https://doi.org/10.1037/h0040957>
- Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications*, 19(4), 497–515.
<https://doi.org/10.1007/s10260-010-0142-z>
- Csépes, I., & Együd, G. (2000). *Into Europe – Prepare for modern English exams: The speaking handbook*. Budapest: Teleki László Foundation; British Council.

https://www.lancaster.ac.uk/fass/projects/examreform/into_europe/speaking.pdf

Cui, J.-F., Wang, Y., Shi, H.-S., Lui, L.-L., Chen, X.-J., Chen, Y.-H. (2015).

Effects of working memory load on uncertain decision-making: evidence from the Iowa Gambling Task. *Frontiers in Psychology*, 6.

<https://doi.org/10.3389/fpsyg.2015.00162>

Cumming, A. (1990). Expertise in evaluating second language compositions.

Language Testing, 7(1), 31–51.

<https://doi.org/10.1177/026553229000700104>

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating

ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, 86(1), 67–96. <https://doi.org/10.1111/1540-4781.00137>

Curşeu, P. L., & Schruijer, S. G. L. (2012). Decision styles and rationality: An

analysis of the predictive validity of the General Decision-Making Style Inventory. *Educational and Psychological Measurement*, 72(6), 1053–1062. <https://doi.org/10.1177/0013164412448066>

Dane, E., Rockmann, K., & Pratt, M.E. (2012). When should I trust my gut?

Linking domain expertise to intuitive decision-making effectiveness. *Organizational Behavior and Human Decision Processes*, 119(2), 187–194. <https://doi.org/10.1016/j.obhdp.2012.07.009>

Davidson, F. (2000). Book Review: Standards for educational and psychological testing. *Language Testing*, 17(4), 457–462.

<https://doi.org/10.1177/026553220001700405>

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999).

Dictionary of language testing. Cambridge: Cambridge University Press.

- Davis, L. (2008). *Rater expertise in a second language speaking assessment: The influence of training and experience* (Publication No. 3569056) [Doctoral dissertation, University of Hawai'i at Manoa]. ProQuest Dissertations Publishing.
- Davis, L. (2015). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, 33(1), 117–135.
<https://doi.org/10.1177/0265532215582282>
- Davis, L. (2018). Analytic, Holistic, and Primary Trait Marking Scales. In *The TESOL Encyclopaedia of English Language Teaching* (pp. 1–6). Hoboken, NJ, USA: John Wiley & Sons.
- Delaney, R., Strough, J., Parker, A. M., & de Bruin, W. B. (2015). Variations in decision-making profiles by age and gender: A cluster-analytic approach. *Personality and Individual Differences*, 85, 19–24.
<https://doi.org/10.1016/j.paid.2015.04.034>
- De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, 34(1), 5–34. <https://doi.org/10.1017/S0272263111000489>
- De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2013). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, 36(2), 223–243.
<https://doi.org/10.1017/s0142716413000210>
- Denzin, N. K. (2012). Triangulation 2.0. *Journal of Mixed Methods Research*, 6(2), 80–88. <https://doi.org/10.1177/1558689812437186>
- Dörnyei, Z. (2007). *Research methods in applied linguistics*. New York: Oxford.

- Duff, P. A. (2019) Case study research: Making language learning complexities visible. In J. McKinley & H. Rose (Eds.), *The Routledge Handbook of Research Methods in Applied Linguistics* (pp. 144-153). Taylor & Francis Group.
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155–185.
<https://doi.org/10.1177/0265532207086780>
- Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.
<https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000001680667a23>
- Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, 9(3), 270–292. <https://doi.org/10.1080/15434303.2011.649381>
- Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement* (Second edition). Frankfurt am Main: Peter Lang.
- Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, 33(1), 56–70.
<https://doi.org/10.1111/j.1745-3984.1996.tb00479.x>
- Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales*. New York, London: Routledge.

- Engelhard, G., Jr, Wang, J., & Wind, S. A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling*, 60(1), 33–52.
- Epstein, S. (1991). Cognitive-experiential self-theory: An integrative theory of personality. In R. C. Curtis (Ed.), *The relational self: Theoretical convergences in psychoanalysis and social psychology* (pp. 111–137). The Guilford Press.
- Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, 49(8), 709–724.
<https://doi.org/10.1037/0003-066X.49.8.709>
- Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, 71(2), 390–405.
<https://doi.org/10.1037/0022-3514.71.2.390>
- Ercikan K., & Pellegrino, J. W. (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. New York, NY: Taylor & Francis.
- Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3), 363–406.
- Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation Framework. *Frontiers in Psychology*, 11.
<https://doi.org/10.3389/fpsyg.2020.00330>

- Field, A. P. (2014). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). London: Sage.
- Fischer, S., Soye, K., & Gurtner, S. (2015). Adapting Scott and Bruce's General Decision-Making Style Inventory to patient decision making in provider choice. *Medical Decision Making*, 35(4), 525–532.
<https://doi.org/10.1177/0272989x15575518>
- Fletcher, J. M., Marks, A. D. G., & Hine, D. W. (2012). Latent profile analysis of working memory capacity and thinking styles in adults and adolescents. *Journal of Research in Personality*, 46(1), 40–48.
<https://doi.org/10.1016/j.jrp.2011.11.003>
- Fulcher G. (2017) Criteria for evaluating language quality. In Shohamy E., Or I., & May S. (Eds.) *Language Testing and Assessment. Encyclopedia of Language and Education* (third edition). Springer, Cham.
https://doi.org/10.1007/978-3-319-02261-1_13
- Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Education.
- Fulcher, G. (2015). *Re-examining Language Testing*. Oxon & New York: Routledge.
- Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28(1), 5–29. <https://doi.org/10.1177/0265532209359514>
- Gaetano, J. (2018). Holm-Bonferroni sequential correction: An Excel calculator (1.3) [Microsoft Excel workbook].
https://www.researchgate.net/publication/322568540_Holm-Bonferroni_sequential_correction_An_Excel_calculator_13

- Galotti, K. M., & Tinkelenberg, C. E. (2009). Real-life decision making: Parents choosing a first-grade placement. *American Journal of Psychology*, 122(4), 455–468. <http://www.jstor.org/stable/27784421>
- Gambetti, E., Fabbri, M., Bensi, L., & Tonetti, L. (2008). A contribution to the Italian validation of the General Decision-making Style Inventory. *Personality and Individual Differences*, 44(4), 842–852. <https://doi.org/10.1016/j.paid.2007.10.017>
- Gilbert, S., & Staub, G. (2014). Examiner confidence survey: An investigation into Speaking Examiners' confidence in the accuracy of the assessments they make. *Cambridge English Research Note*, 57, 50–59.
- Glöckner, A. & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgement and Decision Making*, 3(3), 215–228.
- Green, R. (2013). *Statistical analyses for language testers*. Basingstoke: Palgrave Macmillan.
- Guest, G., MacQueen, K. M. & Namey, E. E. (2012). Introduction to applied thematic analysis. In G. Guest, K. M. MacQueen, & E. E. Namey (Eds.) *Applied thematic analysis* (pp. 3–20). Thousand Oaks, CA: SAGE Publications, Inc. <https://doi.org/10.4135/9781483384436>
- Gunnell, J. J., & Ceci, S. J. (2010). When emotionality trumps reason: A study of individual processing style and juror bias. *Behavioral Sciences & the Law*, 28(6), 850–877. <https://doi.org/10.1002/bsl.939>
- Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability* (fourth edition). Gaithersburg: Advanced Analytics, LLC.

- Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.) *Assessing Second Language Writing in Academic Contexts* (pp. 241–76). Norwood, NJ: Ablex.
- Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, 12(1), 1–9.
<http://doi.org/10.1016/j.asw.2007.05.002>
- Handley, S. J., Newstead, S. E., & Wright, H. (2000). Rational and experiential thinking: A study of the REI. In R. J. Riding & S. G. Rayner (Eds.) *International perspectives on individual differences* (pp. 97–113). Stamford, CO: Ablex.
- Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, 20(3), 281–307.
<https://doi.org/10.1080/0969594x.2012.742422>
- Holm, S. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Holzknicht, F., Kremmel, B., Konzett, C., Eberharter, K., & Spöttl, C. (2018). Potentials and challenges of teacher involvement in rating scale design for high-stakes exams. In D. Xerri & P. Vella Briffa (Eds.), *Teacher involvement in high-stakes testing* (pp. 1–20).
- Horstmann, N., Ahlgrimm, A., & Glöckner, A. (2009). How distinct are intuition and deliberation? An eye-tracking analysis of instruction-induced decision modes. *Judgement and Decision Making*, 4(5), 335–354.
- Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System*, 41(3), 770–785. <https://doi.org/10.1016/j.system.2013.07.009>

- Huang, B., Alegre, A., & Eisenberg, A. (2016). A Cross-Linguistic Investigation of the Effect of Raters' Accent Familiarity on Speaking Assessment. *Language Assessment Quarterly*, 13(1), 25–41.
<https://doi.org/10.1080/15434303.2015.1134540>
- Hughes, A. (2003). *Testing for Language Teachers* (second edition). Cambridge University Press: Cambridge.
- Huhta, A., Ohranen, S., Halonen, M., Hirvelä, T., Neittaanmäki, R., Ahola, S., Ullakonoja, R. (2019, March). Rater behaviour in a high-stakes L2 examination: Does a test taker's perceived first language matter? [Paper presentation]. 41st Language Testing Research Colloquium, Atlanta, GA.
- Indrarathne, B., & Kormos, J. (2017). The role of working memory in processing L2 input: Insights. *Bilingualism: Language and Cognition*, 21(2), 355–374. <https://doi.org/10.1017/s1366728917000098>
- Isaacs, T., & Trofimovich, P. (2010). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, 32(1), 113–140.
<https://doi.org/10.1017/s0142716410000317>
- Ivankova, N. V., & Greer, J. L. (2015). Mixed methods research and analysis. In B. Paltridge & A. Phakiti (Eds.), *Research Methods in Applied Linguistics* (pp. 63–81). Bloomsbury Publishing.
- Jacquemot, C., & Scott, S. K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences* 10, 480–486. <https://doi.org/10.1016/j.tics.2006.09.002>
- Jahnukainen, M. (2010). Extreme cases. In A. J. Mills, G. Durepos & E. Wiebe (Eds.), *Encyclopedia of case study research* (pp. 379–380), Thousand

Oaks, CA: SAGE Publications, Inc.

<https://doi.org/10.4135/9781412957397.n142>

- Jang, E. E., Wagner, M., & Park, G. (2014). Mixed Methods Research in Language Testing and Assessment. *Annual Review of Applied Linguistics*, 34, 123–153. <https://doi.org/10.1017/s0267190514000063>
- Joffe, H. (2012). Thematic analysis. In D. Harper & A. R. Thompson (Eds.), *Qualitative research methods in mental health and psychotherapy* (pp. 209–223), John Wiley & Sons, Ltd.
<https://doi.org/10.1002/9781119973249.ch15>
- Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioural economics. *American Economic Review*, 93(5), 1449–1475.
<https://doi.org/10.1257/000282803322655392>
- Kahneman, D. & Frederick, S. (2005). A model of heuristic judgement. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 267–293). Cambridge: Cambridge University Press.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535. <https://doi.org/10.1037/0033-2909.112.3.527>
- Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspective*, 2(3), 135–170. https://doi.org/10.1207/s15366359mea0203_1
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (fourth edition, pp. 17–64). Westport: American Council on Education and Praeger.

- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
<https://doi.org/10.1111/jedem.12000>
- Kane, M., Crooks, T., & Cohen, A. D. (1999). Validating measures of performance. *Educational Measurement Issues and Practice*, 18(2), 5–17.
<https://doi.org/10.1111/j.1745-3992.1999.tb00010.x>
- Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly*, 9(3), 249–269.
<https://doi.org/10.1080/15434303.2011.642631>
- Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, 36(4), 481–504. <https://doi.org/10.1177/0265532219849522>
- Karlsson, L., Juslin, P., & Olsson, H. (2008). Exemplar-based inference in multi-attribute decision making: Contingent, not automatic, strategy shifts? *Judgement and Decision Making*, 3(3), 244–260.
<http://journal.sjdm.org/bn5/bn5.html>
- Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment* (Publication No. 3448033) [Doctoral dissertation, Teachers College, Columbia University]. ProQuest Dissertations Publishing.
- Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, 12(3), 239–261.
<https://doi.org/10.1080/15434303.2015.1049353>

- Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, 26(2), 187–217.
<https://doi.org/10.1177/0265532208101010>
- Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, 26(2), 275–304.
<https://doi.org/10.1177/0265532208101008>
- Knoch, U., & Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. *Language Testing*, 4(2), 1–23.
<https://doi.org/10.1177/0265532217710049>
- Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19(1), 3–31. <https://doi.org/10.1191/0265532202lt218oa>
- Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, 11(2), 261–271.
<https://doi.org/10.1017/s1366728908003416>
- LaFlair, G. T., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, *t*-tests and ANOVAs. In L. Plonsky (Ed.) *Advancing Quantitative Methods in Second Language Research* (pp. 46–77). Routledge. <https://doi.org/10.4324/9781315870908>
- Läge, D. & Hausmann, D. (2008). Sequential evidence accumulation in decision making: The individual desired level of confidence can explain the extent of information acquisition. *Judgement and Decision Making*, 3(3), 229–243.

- LeBoeuf, Robyn A. & Shafir, Eldar B. (2005). Decision Making. In K. J. Holyoak, & R. G. Morrison (Eds.) *The Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press. Chapter 11, 243–266.
- Leighton, J. P. (2012). Editorial. *Educational measurement: Issues and practice*, 31(3), 1.
- Li, Z., Banerjee, J., & Zumbo, B. D. (2017). Response time data as validity evidence: Has it lived up to its promise and, if not, what would it take to do so. In B. D. Zumbo & A. M. Hubley (Eds.), *Social indicators research series: Vol. 69. Understanding and investigating response processes in validation research* (pp. 159–177). Springer International Publishing.
https://doi.org/10.1007/978-3-319-56129-5_9
- Lim, G. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, 28(4), 543–560.
<https://doi.org/10.1177/0265532211406422>
- Linacre, J. M. (1994). *Many-facet Rasch measurement* (Second Edition, pp. 1–158). Chicago: Mesa Press.
- Linacre, J. M. (2003). Rasch power analysis: Size vs. significance: Infit and outfit standardized chi-square fit statistic. *Rasch Measurement Transactions*, 17(1), 918.
- Linacre, J. M. (2020). *A user's guide to FACETS: Rasch-model computer programs*. Chicago: Winsteps.com. <https://www.winsteps.com/a/Facets-Manual.pdf>

- Loo, R. (2000). A psychometric evaluation of the General Decision-Making Style Inventory. *Personality and Individual Differences*, 29(5), 895–905.
[https://doi.org/10.1016/s0191-8869\(99\)00241-x](https://doi.org/10.1016/s0191-8869(99)00241-x)
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, 19(3), 246–276.
<https://doi.org/10.1191/0265532202lt230oa>
- Lumley, T. (2005). Assessing second language writing: The rater's perspective. Frankfurt: Lang.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54–71.
<https://doi.org/10.1177/026553229501200104>
- Mackey, A., & Gass, S. M. (2015). Mixed methods. In *Second language research* (second edition, pp. 275–291). Taylor & Francis Group.
- Martins, A. (2016). A review of important cognitive concepts in aviation. *Aviation*, 20(2), 65–84.
<https://doi.org/10.3846/16487788.2016.1196559>
- May, L. A. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing (MPLT)*, 11(1), 29–51.
- McLaughlin, J. E., Cox, W. C., Williams, C. R., & Shepherd, G. (2014). Rational and experiential decision-making preferences of third-year student pharmacists. *American Journal of Pharmaceutical Education*, 78(6), 120.
<https://doi.org/10.5688/ajpe786120>
- McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

- Messick S. (1989). Validity. In: Linn, R. L. (Ed.), *Educational Measurement* (pp. 13–103). New York: American Council on Education and Macmillan.
- Miyahara, M. (2019). Sampling: Problematizing the issue. In J. McKinley & H. Rose (Eds.), *The Routledge handbook of research methods in applied linguistics* (pp. 52–62).
<https://www.taylorfrancis.com/chapters/edit/10.4324/9780367824471-5/sampling-masuko-miyahara>
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “Frontal Lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49–100.
<https://doi.org/10.1006/cogp.1999.0734>
- Murphy, K. R., Jako, R. A., and Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology*, 78(2), 218–225. <https://doi.org/10.1037/0021-9010.78.2.218>
- Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement Issues and Practice*, 31(3), 48–49.
<http://doi.org/10.1111/j.1745-3992.2012.00243.x>
- Myford, C. M., & Wolfe, E. W. (2000). Monitoring sources of variability within the Test of Spoken English Assessment System. *TOEFL Research Report*, 65. <https://www.ets.org/Media/Research/pdf/RR-00-06-Myford.pdf>
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386–422.

- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189–227.
- Nakatsuhara, F., Inoue C., & Taylor, L. (2020). Comparing rating modes: Analysing live, audio, and video ratings of IELTS Speaking Test performances, *Language Assessment Quarterly*, 18(2), 83–106. <https://doi.org/10.1080/15434303.2020.1799222>
- Newell, B. R., & Bröder, A. (2008). Cognitive processes, models and metaphors in decision research. *Judgement and Decision Making*, 3(3), 195–204.
- Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. *System*, 30(2), 143–154. [https://doi.org/10.1016/s0346-251x\(02\)00002-7](https://doi.org/10.1016/s0346-251x(02)00002-7)
- Österreichisches Sprachen-Kompetenz-Zentrum (Hrsg.). (2020). *Mündliche Reifeprüfung Englisch. Modellaufgaben und Videoperformanzen auf dem Niveau B2*. Graz: ÖSZ. http://www.oesz.at/download/publikationen/muendlichereifepuefungenglisch_modellaufgabenniveaub2_web.pdf
- Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, 76(6), 972–987. <https://doi.org/10.1037//0022-3514.76.6.972>
- Park, M. S. (2020). Rater effects on L2 oral assessment: Focusing on accent familiarity of L2 teachers. *Language Assessment Quarterly*, 17(3), 231–243. <https://doi.org/10.1080/15434303.2020.1731752>

- Payne, J. W., & Venkatraman, V. (2011). Opening the black box: conclusions to A handbook of process tracing methods for decision research. In M. Schulte-Mecklenbeck, A. Kühberger, & R. Ranyard (Eds.), *A handbook of process tracing methods for decision research* (pp. 223–249). New York: Psychology Press.
- Phillips, W. J., Fletcher, J. M., Marks, A. D. G., & Hine, D. W. (2015). Thinking Styles and Decision Making: A Meta-Analysis. *Psychological Bulletin*, 1–32. <https://doi.org/10.1037/bul0000027>
- Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 74–91). Cambridge, England: Cambridge University Press.
- Purpura, J. E. (2013). Cognition and language assessment. In A. J. Kunnan (Ed.), (First, pp. 1–25). John Wiley & Sons, Inc.
<https://doi.org/10.1002/9781118411360.wbcla150>
- Raczynski, K. R., Cohen, A. S., Engelhard, G., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative Frame-of-Reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, 52(3), 301–318.
<https://doi.org/10.1111/jedm.12079>
- Reisen, N., Hoffrage, U., & Mast, F. (2008). Identifying decision strategies in a consumer choice situation. *Judgment and Decision Making*, 3, 641–658.
- Révész, A., Michel, M., & Lee, M. J. (2017). *Investigating IELTS Academic Writing Task 2: Relationships between cognitive writing processes, text*

quality, and working memory. IELTS Research Report, 2017/3. London:
The British Council.

- Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, 124(2), 207–231. <https://doi.org/10.1037/0096-3445.124.2.207>
- Russo, E.J., & Doshier, B.A. (1983). Strategies for multiattribute binary choice. *Journal of Personality and Social Psychology: Learning, Memory and Cognition*, 59(4), 676–696. <https://doi.org/10.1037/0278-7393.9.4.676>
- Salkind, N. J. (2010). *Encyclopedia of research design* (Vols. 1-0). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781412961288
- Sánchez, E., Fernández-Berrocal, P., Alonso, D., & Tuma, E. (2012). Measuring both systems of reasoning: a study of the predictive capacity of a new version of the Rational-Experiential Inventory. *European Journal of Education and Psychology*, 5(2), 121–132. <https://doi.org/10.1989/ejep.v5i2.96>
- Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, 24(3), 355–390. doi:10.1177/0265532207077205
- Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465–493. doi: 10.1177/0265532208094273
- Schellig, D., Drechsler, R., Heinemann, D., & Sturm, W. (Eds.) (2009). *Handbuch neuropsychologischer Testverfahren – Handbook of Neuropsychological Tests – Attention, Memory, Executive Functions*. Göttingen: Hogrefe Verlag.

- Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (Eds.) (2011). *A handbook of process tracing methods for decision research: A critical review and user's guide*. New York and Hove: Taylor & Francis.
- Scott, S. G., & Bruce, R. A. (1995). Decision-Making Style: The development and assessment of a new measure. *Educational and Psychological Measurement*, 55(5), 818–831.
<https://doi.org/10.1177/0013164495055005017>
- Shepard, L. A. (1993). Evaluating Test Validity. *Review of Research in Education*, 19, 405–450. <https://doi.org/10.2307/1167347>
- Shirzadifard, M., Shahghasemi, E., Hejazi, E., Naghsh, Z., & Ranjbar, G. (2018). Psychometric properties of Rational-Experiential Inventory for Adolescents. *SAGE Open*, 8(1).
<https://doi.org/10.1177/2158244018767219>
- Simon, D. (2004). A third view of the black box: Cognitive coherence in legal decision making. *The University of Chicago Law Review*, 71(2), 511–586.
- Sireci, S. G. (2016). On validity theory and test validation. *Educational Researcher*, 36(8), 477–481. <https://doi.org/10.3102/0013189x07311609>
- Spicer, D. P., & Sadler-Smith, E. (2005). An examination of the general decision making style questionnaire in two UK samples. *Journal of Managerial Psychology*, 20(2), 137–149. <https://doi.org/10.1108/02683940510579777>
- Spiro, R. J., & Jehng, J.-C. (1990). Cognitive flexibility and hypertext: Theory and technology for the nonlinear and multidimensional traversal of complex subject matter. In D. Nix & R. J. Spiro (Eds.), *Cognition, education, and multimedia: Exploring ideas in high technology* (p. 163–205). Lawrence Erlbaum Associates, Inc.

- Spiro, R., Collins, B., Thota, J., & Feltovich, P. (2003). Cognitive flexibility theory: Hypermedia for complex learning, adaptive knowledge application, and experience acceleration. *Educational Technology*, 43(5), 5–10. <http://www.jstor.org/stable/44429454>
- Spiro, R.J., Coulson, R.L., Feltovich, P.J., & Anderson, D. (1988). Cognitive flexibility theory: Advanced knowledge acquisition in ill-structured domains. In V. Patel (ed.), *Proceedings of the 10th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Spöttl, C., Kremmel, B., Holzknecht, F., & Alderson, J. C. (2016). Evaluating the achievements and challenges in reforming a national language exam: The reform team's perspective. *Melbourne Papers in Language Testing (MPLT)*, 5(1), 1–22.
- St-Onge, C., Chamberland, M., Lévesque, A. (2016). Expectations, observations, and the cognitive processes that bind them: expert assessment of examinee performance. *Advances in Health Sciences Education*, 21(3), 627–642. <https://doi.org/10.1007/s10459-015-9656-3>
- Stake, R. E. (2005). Qualitative case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (p. 443–466). Sage Publications Ltd.
- Stake, R. E. (2006). *Multiple case study analysis*. New York, NY: Guilford.
- Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation*, 9(4), 1–11.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680. <http://www.jstor.org/stable/1671815>

- Stone, M. H., & Wright, B. D. (1979). *Best test design*. Chicago, IL: Mesa.
- Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educational Measurement: Issues and Practice*, 31(3), 21–30. <https://doi.org/10.1111/j.1745-3992.2012.00240.x>
- Suto, I. & Greateorex, J. (2008). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policy & Practice*, 15(1), 73–89. <https://doi.org/10.1080/09695940701876177>
- Taylor, L. & Galaczi, E. (2011) Scoring validity. In L. Taylor (Ed.), *Examining Speaking: Research and practice in assessing second language speaking* (pp. 171–233). Cambridge: Cambridge University Press.
- Teddle, C., & Tashakkori, A. (2009). *Foundations of mixed methods research*. Thousand Oaks, CA: SAGE.
- Thunholm, P. (2004). Decision-making style: Habit, style or both? *Personality and Individual Differences*, 36(4), 931–944. [https://doi.org/10.1016/s0191-8869\(03\)00162-4](https://doi.org/10.1016/s0191-8869(03)00162-4)
- Thunholm, P. (2008). Decision-making styles and physiological correlates of negative stress: Is there a relation? *Scandinavian Journal of Psychology*, 49(3), 213–219. <https://doi.org/10.1111/j.1467-9450.2008.00640.x>
- Tombaugh, T. (2004). Trail Making Test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology*, 19(2), 203–214. [https://doi.org/10.1016/s0887-6177\(03\)00039-8](https://doi.org/10.1016/s0887-6177(03)00039-8)
- Tonetti, L., Fabbri, M., Boreggiani, M., Guastella, P., Martoni, M., Ruiz Herrera, N., & Natale, V. (2016). Circadian preference and decision-making

styles. *Biological Rhythm Research*, 47(4), 573–581.

<https://doi.org/10.1080/09291016.2016.1167312>

Toulmin, S. (2003). *The uses of argument* (second edition). Cambridge:

Cambridge University Press. <https://doi.org/10.1017/cbo9780511840005>

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge: Cambridge University Press.

Upton, G., & Cook, I. (2014). Cohen's kappa. In G. Upton & I. Cook (Eds.), *A dictionary of statistics* (third edition). Oxford: Oxford University Press.

VERBI Software. (2017). MAXQDA 2018 [computer software]. Berlin, Germany: VERBI Software. Available from maxqda.com.

Wagner, E. (2015). Survey research. In B. Paltridge & A. Phakiti (Eds.), *Research methods in Applied Linguistics: A practical resource* (pp. 83–100). London: Bloomsbury.

Wallace, M. P. (2020). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*. <https://doi.org/10.1111/lang.12424>

Wang, J. & Engelhard, G. (2017). Using a multifocal lens model and Rasch measurement theory to evaluate rating quality in writing assessments. *Pensamiento Educativo: Journal of Latin-American Educational Research*, 54(2), 1–16.

Wang, J., & Luo, K. (2019). Evaluating rater judgments on ETIC Advanced writing tasks: An application of generalizability theory and many-facets Rasch model. *Papers in Language Testing and Assessment*, 8(2) 91–116.

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT Speaking Tasks. *Language*

Assessment Quarterly, 12(3), 283–304.

<https://doi.org/10.1080/15434303.2015.1037446>

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language*

Testing, 15(2), 263–287. <https://doi.org/10.1177/026553229801500205>

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press: Cambridge.

Weir, C. J. (2005). *Language testing and validation*. London: Palgrave Macmillan UK.

White, E., McMahon, M., Walsh, M., Coffey, J., & O'Sullivan, L. (2017). Toward a Model of Human Information Processing for Decision-Making and Skill

Acquisition in Laparoscopic Colorectal Surgery. *Journal of Surgical*

Education, 75. <https://doi.org/10.1016/j.jsurg.2017.09.010>

Wickens, C. D. (1992). *Engineering psychology and human performance* (second edition). New York: Harper Collins.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305–319. <https://doi.org/10.1177/026553229301000306>

Wind, S. A., & Peterson, M. E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, 4(4).

<https://doi.org/10.1177/0265532216686999>

Winke, P., & Gass, S. (2012). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation.

TESOL Quarterly, 47(4), 762–789. <http://www.jstor.org/stable/43267928>

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231–

252. <https://doi.org/10.1177/0265532212456968>

- Winke, P., & Lim, G. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, 25, 38–54. <https://doi.org/10.1016/j.asw.2015.05.002>
- Wiseman, C. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, 17(3), 150–173. <https://doi.org/10.1016/j.asw.2011.12.001>
- Witteman, C., Van den Bercken, J., Claes, L., & Godoy, A. (2009). Assessing rational and intuitive thinking styles. *European Journal of Psychological Assessment*, 25(1), 39–47. <https://doi.org/10.1027/1015-5759.25.1.39>
- Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, 4(1), 83–106. [https://doi.org/10.1016/s1075-2935\(97\)80006-2](https://doi.org/10.1016/s1075-2935(97)80006-2)
- Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, 46(1), 35–51.
- Wolfe, E. W. (2006). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment*, 2(1), 37–56.
- Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement Issues and Practice*, 31(3), 31–37. <https://doi.org/10.1111/j.1745-3992.2012.00241.x>
- Wolfe, E. W., Kao, C.-W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, 15(4), 465–492. <https://doi.org/10.1177/0741088398015004002>

- Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*.
media.metrik.de/uploads/incoming/.../RaschMeasurement_Complete.pdf
- Xi, X. (2008). Methods of test validation. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education* (Vol. 7, pp. 177–196). Boston, MA: Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30424-3_175
- Xi, X., & Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT™ Speaking Section and what kind of training helps?* (TOEFLiBT-11). New Jersey: ETS, Princeton.
- Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, 31(4), 501–527. <https://doi.org/10.1177/0265532214536171>
- Yntema, D. B. (1963). Keeping track of several things at once. *Human Factors*, 5(1), 7–17. <https://doi.org/10.1177/001872086300500102>
- Youn, S. J. (2018). Rater variability across examinees and rating criteria in paired speaking assessment. *Papers in Language Testing and Assessment*, 7(1), 32–60.
- Zander, T., Horr, N. K., Bolte, A., & Volz, K. G. (2016). Intuitive decision making as a gradual process: investigating semantic intuition-based and priming-based decisions with fMRI. *Brain and Behavior*, 6(1). <https://doi.org/10.1002/brb3.420>
- Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing*, 27, 37–53. <https://doi.org/10.1016/j.asw.2015.11.001>

- Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, 28(1), 31–50.
<https://doi.org/10.1177/0265532209360671>
- Zhang, Y., & Elder, C. (2013). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy & Practice*, 21(3), 306–325.
<https://doi.org/10.1080/0969594X.2013.845547>
- Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Social indicators research series: Vol. 69. Understanding and investigating response processes in validation research*. Springer International Publishing AG.
<https://doi.org/10.1007/978-3-319-56129-5>

Appendix A Variable Operationalisations

Table Appendix 1 Overview of dependent and independent variables

Variable family	Variable Name	Definition	Measurement	Variable Type	Measurement level
Rating quality (based on scoring patterns)	Severity (Full model)	Severity estimate based on RSM model including all data, higher measure = higher severity	MFRM, Logit	Dependent	Continuous
	Severity (TA), Severity (FLIN), Severity (RSL), Severity (ASL)	Severity estimate based on criterion-specific RSM model, higher measure = higher severity	MFRM, Logit	Dependent	Continuous
	Accuracy (Full model)	Accuracy estimate based on RAM model including all data, higher measure = higher accuracy	MFRM, Logit	Dependent	Continuous
	Accuracy (TA), Accuracy (FLIN), Accuracy (RSL), Accuracy (ASL)	Accuracy estimate based on criterion-specific RAM model, higher measure = higher accuracy	MFRM, Logit	Dependent	Continuous
Rating behaviour metrics (based on time stamp data)	Deliberation time (DT)	How much time did a participant spend on a performance in Session 1 and 2?	Mean duration (s), square root transformed	Dependent	Continuous
	Time to first decision (TTFD)	When did a participant enter their first decision in Session 1 and 2?	Mean duration (s) relative to starting rating a performance	Dependent	Continuous
	Revision count (RC)	How many times did a participant revise their ratings in Session 1 and 2?	Count, square root transformed	Dependent	Continuous
Cognitive processing	Stroop Effect	How well do participants inhibit incorrect response?	Time (ms), reflected	Independent	Continuous

(based on cognitive test scores)	Numbers-Letters Task	How well do participants switch between two tasks?	Time (ms), reflected	Independent	Continuous
	Digit Span Task	What is the participants' phonological short-term memory capacity?	Count	Independent	Continuous
	Keep Track Task	How well can participants manipulate information stored in WM?	Count	Independent	Continuous
	Trail Making Task	How much slower are participants on divided attention task?	Time (ms), reflected	Independent	Continuous
Cognitive processing mode (based on REI-40 questionnaire)	Rationality (Ability, Engagement)	How much does the participant prefer rational processing?	Mean score	Independent	Continuous
	Experientiality (Ability, Engagement)	How much does the participant prefer experiential processing?	Mean score	Independent	Continuous
	Processor type	Do raters have an overriding preferred cognitive processing mode (Rational or experiential)?	Binary score	Independent	Categorical
Decision-making style (based on GDMSI questionnaire)	Rational, Intuition, Dependent, Avoidant, Spontaneous	How much do raters prefer particular decision-making styles according to the GDMSI	Mean score	Independent	Continuous

Appendix B Rating Materials

Appendix B.1 Analytic rating scale

Band	Task Achievement [TA]	Fluency & Interaction [FLIN]	Range of Spoken Language [RSL]	Accuracy of Spoken Language [ASL]
10	(1) All aspects of the task addressed and convincingly expanded (2) Very clear, systematically developed descriptions and presentations, effective highlighting of significant points (3) Accounts for and sustains opinions convincingly	(1) High degree of fluency and spontaneity (2) Intervenes appropriately, frequently relating her/his own contribution to those of others (3) Easily adjusts to level of formality (4) Remarkable ease of expression in longer complex stretches of speech is consistent	(1) Expresses her/himself very clearly, no restriction (2) Very wide range of vocabulary for the task (3) Seldom needs to use circumlocution or paraphrase (4) Uses a wide range of complex structures	(1) Lexical accuracy very high, hardly any incorrect word choice (2) Very good grammatical control (3) Hardly any lexical or grammatical slips (4) Clear, natural pronunciation; uses intonation appropriately to highlight significant points
9				
8	(1) All aspects of the task addressed and expanded (2) Clear, systematically developed descriptions and presentations, appropriate highlighting of significant points (3) Accounts for and sustains opinions well by providing relevant support	(1) Remarkable fluency and spontaneity (2) Frequently intervenes appropriately in discussion (3) Can adjust to level of formality (4) Remarkable ease of expression in even longer complex stretches of speech	(1) Expresses her/himself clearly without much restriction (2) Wide range of vocabulary for the task, varies formulation to avoid repetition (3) Can use circumlocution and paraphrase with ease (4) Uses a range of complex structures	(1) Lexical accuracy high, occasional slips do not hinder communication (2) Good grammatical control, slips or non-systematic errors are rare (3) Slips and errors often corrected in retrospect (4) Clear, natural pronunciation; uses intonation appropriately to highlight significant points
7				
6	(1) Most aspects of the task addressed and sufficiently expanded (2) Clear, detailed descriptions and presentations, expanding and supporting ideas with subsidiary points (3) Accounts for and sustains opinions by providing relevant support	(1) Fluent and spontaneous performance, causing no strain on the listener (2) Effective turntaking, not always elegant (3) Adjusts to changes of direction in conversation (4) Produces stretches of language with a fairly even tempo; few noticeably long pauses	(1) Sufficient range of language for the task, some restriction (2) Good range of vocabulary for the task, varies formulation to avoid frequent repetition (3) Can use circumlocution and paraphrase (4) Uses some complex structures	(1) Lexical accuracy generally high, mistakes do not hinder communication (2) Grammatical control relatively high; any mistakes do not cause misunderstanding (3) Can correct slips and errors if she/he becomes conscious of them (4) Clear, natural pronunciation and intonation
5				
4	(1) Only some aspects of the task addressed but not sufficiently expanded (2) Descriptions and presentations lack clarity and detail (3) Seldom accounts for and sustains opinions	(1) Performance imposes strain on the listener due to lack of fluency and spontaneity (2) Has difficulty intervening in a discussion, turntaking not effective (3) Has difficulty adjusting to changes of direction (4) Frequent stretches of language with uneven tempo; frequent hesitation, some non-productive pauses	(1) Insufficient range of language for parts of the task, frequent restrictions (2) Limited range of vocabulary for the task, lack of range causes repetition (3) Has difficulty using circumlocution or paraphrase (4) Hardly any complex structures	(1) Insufficient lexical and grammatical control (2) Accuracy influenced by L1, errors frequently impede communication (3) Fails to correct mistakes which have caused misunderstandings (4) Pronunciation not always natural, mispronunciations
3				
2	(1) Only some aspects of the task addressed, none expanded (2) Descriptions only presented as a linear sequence of points (3) Fails to account for and sustain opinions (4) Fails to produce sustained language performance	(1) Performance imposes considerable strain on the listener due to lack of fluency and spontaneity (2) Fails to intervene appropriately, little evidence of turntaking (3) Fails to adjust to changes of direction (4) Uneven tempo; frequent hesitation with non-productive pauses	(1) Insufficient range of language for the task (2) Insufficient vocabulary for the task (3) Fails to cover linguistic gaps, foreignises words from L1 (4) No complex structures	(1) Vocabulary elementary; major errors occur when expressing more complex thoughts (2) Accuracy influenced by L1, breakdown of communication (3) Inability to monitor mistakes (4) Accent and/or frequent mispronunciations impede communication
1				
0	(1) Task ignored (2) Not enough language for assessment	(1) Performance hesitant and incoherent throughout (2) Fails to intervene	(1) Not enough language for assessment	(1) Not enough language for assessment

© BMBF & UIBK, 2012

Appendix B.2 Holistic rating forms

II.	Holistic Scale for Interlocutor – B2
------------	---

Candidate:		Class/School:	
Topic/Task:		Date:	

Band			monol. ✓	dial. ✓
10	TA	All aspects of the task addressed and convincingly expanded		
	FLIN	Communicates and interacts with a high degree of fluency and spontaneity		
	RSL	Expresses her/himself clearly with no sign of having to restrict what she/he wants to say		
	ASL	Lexical and grammatical accuracy is very high, hardly any 'slips'		
9				
8	TA	All aspects of the task addressed and expanded		
	FLIN	Communicates and interacts with remarkable fluency and spontaneity		
	RSL	Expresses her/himself clearly and without much sign of having to restrict what she/he wants to say		
	ASL	Lexical and grammatical accuracy is high, 'slips' or non-systematic errors do not hinder communication and are rare		
7				
6	TA	Most aspects of the task addressed and sufficiently expanded		
	FLIN	Fluent and spontaneous performance, causing no strain on the listener		
	RSL	Sufficient range of language for the task, some restriction		
	ASL	Lexical and grammatical accuracy is generally high, mistakes do not hinder communication		
5				
4	TA	Only some aspects of the task addressed but not sufficiently expanded		
	FLIN	Performance imposes strain on the listener due to lack of fluency and spontaneity		
	RSL	Insufficient range of language for parts of the task; frequent restrictions		
	ASL	Insufficient degree of lexical and grammatical control; inability to monitor mistakes		
3				
2	TA	Only some aspects of the task addressed, none expanded		
	FLIN	Performance imposes considerable strain on the listener due to lack of fluency and spontaneity		
	RSL	Insufficient range of language for the task; fails to cover linguistic gaps		
	ASL	Lack of lexical and grammatical control frequently leads to breakdown of communication		
1				
0		Task ignored		
		Fails to produce enough language for assessment		

OVERALL BAND	
---------------------	--

BMBF & UIBK, 2012

Appendix B.3 Analytic rating forms

**Zur Verwendung für das Gespräch zwischen
Interlokutor/in und Kandidat/in**

[illegible]

Appendix C Speaking Tasks

Appendix C.1 Individual long turn “Food waste vs food bank”

Vorbereitungszeit¹¹: 3 Minuten

Sprechzeit¹²: 5 Minuten

PICTURE REMOVED BECAUSE OF COPYRIGHT – picture depicted shop assistant throwing away trays of bread into a wastebin
PICTURE REMOVED BECAUSE OF COPYRIGHT – picture showed person handing a bowl of soup to someone standing in a line

France has become the first country in the world to ban supermarkets from throwing away or destroying unsold food. The shops have to donate the food to charities and food banks. In Austria, there is no such law.

Together with some classmates, you have decided to travel to an international summit and present your point of view on the issue.

Prepare a short presentation:

- contrast the two pictures
- explain how families can avoid food waste
- suggest how schools can raise awareness for this issue among young people

Picture 1: <http://www.dietsciencenews.com/main/solving-the-problem-of-wasted-food-in-america/>

Picture 2: <http://www.taz.de/!5170664/>

¹¹ preparation time

¹² expected speaking time

Appendix C.2 Individual long turn “Enjoying music”

Vorbereitungszeit: 3 Minuten

Sprechzeit: 5 Minuten

PICTURE REMOVED BECAUSE OF PICTURE REMOVED BECAUSE OF
COPYRIGHT – picture showed closeup COPYRIGHT – picture showed band
of young man listening to music with performing on stage in front of an audience
headphones

There are many ways in which we can enjoy art and music. A lot of young people go to live concerts, but tickets can be very expensive. One of the world’s leading music magazines, *The Rolling Stone*, is organizing a global webcast event where young people can share their thoughts about experiencing music in today’s world.

Prepare a short talk in which you:

- compare the two pictures
- explain how you prefer to listen to music
- discuss whether young people are still willing to pay for music

Picture 1: <https://www.flickr.com/photos/gimleteyes/7765639296>

Picture 2: https://www.flickr.com/photos/frf_kmeron/6214092296

Appendix C.3 Paired Speaking

Vorbereitungszeit: keine

Sprechzeit: 10 Minuten

CANDIDATE A

Your local council is planning a meeting to revive a neglected piece of land in your neighborhood. You have been asked to represent your school to argue the needs and requirements of 18-year olds in your area to be included. Use the bullet points below. Discuss with your partner and reach an agreement on the three most important points to present at the meeting.

- sports grounds
- community gardens
- children's playground
- dog park
- barbecue area
- fitness trail

CANDIDATE B

Your local council is planning a meeting to revive a neglected piece of land in your neighborhood. You have been asked to represent your school to argue the needs and requirements of 18-year olds in your area to be included. Use the bullet points below. Discuss with your partner and reach an agreement on the three most important points to present at the meeting.

- dog park
- barbecue area
- fitness trail
- arts sculptures
- swimming ponds
- open air stage

Appendix D Rating Plan for Reference Scores

Figure Appendix 1 Fully rotated rating plan

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5		Rater 6		Group1	Group2
Performance	1	2	1	2	1	2	1	2	1	2	1	2		
P01	x		x		x			o		o		o	R1	R4
P02	o		o		o			x		x		x	R2	R5
P03		o		o		o	x		x		x		R3	R6
P04		x		x		x	o		o		o			
P05		o								o		o	R1	R2
P06		x	o		o		o			x		x	R5	R3
P07	x			o		o		o	x		x		R6	R4
P08	o			x		x		x	o		o			
P09		o		o	x		x		x			o	R1	R3
P10		x		x	o		o		o			x	R2	R4
P11	x		x								x		R6	R5
P12	o		o			x		x		x	o			
P13		o		o	x			o	x		x		R1	R3
P14		x		x	o			x	o		o		R2	R5
P15	x		x			o	x			o		o	R4	R6
P16	o		o			x	o			x		x		
P17	x		x			o		o	x			o	R1	R3
P18	x		x						x				R2	R5
P19					o		o				o		R5	R6
P20		o		o	x		x			o	x			
P21	o			x	o		o			x		x	R1	R2
P22	x			o	x		x			o		o	R3	R5
P23		x	o			x		x	o		o		R4	R6
P24		o	x			o		o	x		x			
P25	o			x	o			x	o			x	R1	R2
P26	x			o	x			o	x			o	R3	R4
P27		x	o			x	o			x	o		R5	R6
P28		o	x			o	x			o	x			
P29		x	o			x	o		o			x	R1	R2
P30		o	x			o	x		x			o	R3	R4
P31	o			x	o			x		x	o		R6	R5
P32	x			o	x			o		o	x			
P33		x						x		x			R1	R2
P34		o	x		x			o		o	x		R4	R3
P35	o			x		x	o		o			x	R5	R6
P36	x			o		o	x		x			o		
P37		x	o		o			x	o			x	R1	R2
P38		o	x		x			o	x			o	R4	R3
P39	o			x		x	o			x	o		R6	R5
P40	o			x		x	o			x	o			
P41	x								x		x		R1 R5 R6	R2 R3 R4
Total	40		37		36		37		39		39			

Appendix E Fair Scores for Selected Performances

Appendix E.1 Fair Scores Session 1

Table Appendix 2 Fair scores for performances in Session 1

Speaker	Proficiency Measure	TA		FLIN		RSL		ASL	
		Obs. Avge	Fair Avge*	Obs. Avge	Fair Avge*	Obs. Avge	Fair Avge*	Obs. Avge	Fair Avge*
40	3.98	9.33	10	10.0	10	9.67	10	9.67	10
37	3.47	10.0	10	10.0	10	9.67	10	9.67	10
3	2.56	9.33	9	8.67	9	9.00	9	8.33	9
33	1.84	8.50	9	8.50	8	8.50	9	8.33	8
14	1.71	8.33	9	8.33	8	8.67	8	8.67	8
11	1.42	8.67	8	7.83	8	8.00	8	7.83	8
15	1.33	7.00	8	7.33	8	8.00	8	9.00	8
6	1.14	8.33	8	7.67	8	8.00	8	8.33	8
10	0.96	8.67	8	8.67	8	7.67	8	8.00	8
1	0.82	7.67	8	6.67	7	8.33	8	7.67	7
30	0.59	8.00	7	7.00	7	6.67	7	6.33	7
31	0.32	7.67	7	7.33	7	7.00	7	7.00	7
8	-0.01	7.00	7	6.00	7	6.33	7	6.67	7
21	-0.05	6.67	7	6.67	7	7.00	7	7.00	7
41	-0.47	6.17	7	6.33	6	6.33	6	6.67	6
<i>M</i>	1.31	8.09	8.05	7.80	7.83	7.92	7.93	7.94	7.83

Note. Ratings for the four criteria range from 6 (below band 6 is below B2 level) and 10 (highest level described in the scale). Rounded fair averages appear in boldface whenever they differ from observed average for the same examinee. TA = task achievement. FLIN = fluency and interaction. RSL = range of spoken language. ASL = accuracy of spoken language.

Appendix E.2 Fair Scores Session 2

Table Appendix 3 Fair scores for performances in Session 2

Speaker	Proficiency Measure	TA		FLIN		RSL		ASL	
		Obs. Avge	Fair Avge*	Obs. Avge	Fair Avge*	Obs. Avge	Fair Avge*	Obs. Avge	Fair Avge*
27	5.68	10.0	10	10.0	10	10.0	10	10.0	10
35	3.26	9.33	10	10.0	9	9.67	10	10.0	9
38	2.67	9.33	9	8.33	9	8.67	9	8.00	9
18	2.08	8.17	9	8.83	9	8.67	9	9.00	9
34	1.67	8.33	8	7.67	8	7.00	8	6.67	8
39	1.56	9.00	8	8.33	8	8.00	8	8.00	8
36	1.55	9.00	8	8.00	8	9.00	8	8.67	8
13	1.52	8.33	8	8.00	8	8.33	8	8.00	8
2	1.01	9.00	8	7.67	8	7.67	8	7.67	8
4	0.87	7.00	8	6.33	7	7.00	8	6.67	7
22	0.82	7.67	8	8.00	7	8.00	8	8.00	7
32	0.78	8.67	8	8.33	7	7.67	8	7.67	7
12	0.36	7.67	7	6.33	7	7.00	7	7.00	7
9	0.18	6.67	7	6.00	7	6.00	7	6.33	7
7	-0.34	7.33	7	7.00	6	7.00	7	6.67	6
<i>M</i>	1.58	8.37	8.21	7.92	8.00	7.98	8.10	7.89	7.99

Note. Ratings for the four criteria range from 6 (below band 6 is below B2 level) and 10 (highest level described in the scale). Rounded fair averages appear in boldface whenever they differ from observed average for the same examinee. TA = task achievement. FLIN = fluency and interaction. RSL = range of spoken language. ASL = accuracy of spoken language.

Appendix F Rating Form

Figure Appendix 2 First page of rating form

Participant: _____

Rating Session 1

Overview

Set 1	8 performances
intermission	2 short questions break (optional)
Set 2	7 performances
Closing	6 short questions 4 open questions

Please note:

- * start playing the videos immediately
- * never pause the videos
- * enter your rating decisions as soon as you feel you have made them (you can still change your decisions before pressing the >> button, but not after)
- * you can take smaller breaks after finishing the rating of a performance and pressing the >> button

Figure Appendix 3 Rating forms

Performance: <input type="text"/>		ROUND <input type="text"/>		Rater number: <input type="text"/>	
TA	FLIN	RSL	ASL		
Justification:					
<hr/>					
<hr/>					
<hr/>					
<hr/>					
<hr/>					
<hr/>					

Performance: <input type="text"/>		ROUND <input type="text"/>		Rater number: <input type="text"/>	
TA	FLIN	RSL	ASL		
Justification:					
<hr/>					
<hr/>					
<hr/>					
<hr/>					
<hr/>					
<hr/>					

Performance: <input type="text"/>		ROUND <input type="text"/>		Rater number: <input type="text"/>	
TA	FLIN	RSL	ASL		
Justification:					
<hr/>					
<hr/>					
<hr/>					
<hr/>					
<hr/>					
<hr/>					

Appendix G Intercorrelations Between Measures of Rating Behaviour

Table Appendix 4 Intercorrelations of mean DT between rating session

Variables	1	2	3	4	5
1. Set 1	-				
2. Set 2	.661**	-			
3. Session 1	.951**	.834**	-		
4. Set 3	.585**	.702**	.663**	-	
5. Set 4	.527**	.561**	.553**	.836**	-
6. Session 2	.589**	.686**	.652**	.975**	.917**

Note. ** Correlation is significant at the 0.01 level (2-tailed).

Table Appendix 5 Intercorrelations of mean TTFD between rating session

Variables	1	2	3	4	5
1. Set 1	-				
2. Set 2	.810**	-			
3. Session 1	.942**	.948**	-		
4. Set 3	.805**	.893**	.889**	-	
5. Set 4	.795**	.790**	.819**	.835**	-
6. Session 2	.827**	.881**	.887**	.954**	.952**

Note. ** Correlation is significant at the 0.01 level (2-tailed).

Appendix H Inter correlations between Rater Quality and Rater Behaviour

Table Appendix 6 Spearman correlations between rater severity and mean DT

Variable	Severity (Full model)	Severity (TA)	Severity (FLIN)	Severity (RSL)	Severity (ASL)
MDT	-.18 [-.49, .15]	-.08 [-.40, .27]	-.28 [-.57, .06]	-.19 [-.49, .13]	-.09 [-.40, .22]

Note. Values in square brackets indicate the 95% confidence interval for each correlation. Bootstrap results based on 2000 samples.

Table Appendix 7 Spearman correlations between rater accuracy and mean DT

Variable	Accuracy (Full model)	Accuracy (TA)	Accuracy (FLIN)	Accuracy (RSL)	Accuracy (ASL)
MDT	0.12 [-.22, .44]	-0.01 [-.37, .34]	0.21 [-.11, .50]	0.12 [-.20, .45]	0.10 [-.24, .45]

Note. Values in square brackets indicate the 95% confidence interval for each correlation. Bootstrap results based on 2000 samples.

Table Appendix 8 Spearman correlations between rater severity and mean TTFD

Variable	Severity (Full model)	Severity (TA)	Severity (FLIN)	Severity (RSL)	Severity (ASL)
MTTFD	-0.01 [-.31, .32]	0.07 [-.27, .39]	-0.13 [-.42, .21]	-0.07 [-.35, .25]	0.10 [-.23, .42]

Note. Values in square brackets indicate the 95% confidence interval for each correlation. Bootstrap results based on 2000 samples.

Table Appendix 9 Spearman correlations between rater accuracy and mean TTFD

Variable	Accuracy (Full model)	Accuracy (TA)	Accuracy (FLIN)	Accuracy (RSL)	Accuracy (ASL)
<i>M</i> TTFD	-0.17	-.44**	0.03	-0.04	-0.05
	[-.22, .44]	[-.37, .34]	[-.11, .50]	[-.20, .45]	[-.24, .45]

Note. Values in square brackets indicate the 95% confidence interval for each correlation. Bootstrap results based on 2000 samples. ** Correlation is significant at the 0.01 level (2-tailed).

Table Appendix 10 Pearson correlations between rater severity and RC

Variable	Session 1	Session 2
Severity (Full model)	-.181	-.084
Severity (TA)	-.274	-.153
Severity (FLIN)	-.047	-.012
Severity (RSL)	-.062	.043
Severity (ASL)	-.232	-.141

Table Appendix 11 Pearson correlations between rater accuracy and RC

Variable	Session 1	Session 2
Accuracy (Full model)	.356	.256
Accuracy (TA)	.552**	.488**
Accuracy (FLIN)	.169	.183
Accuracy (RSL)	.155	.037
Accuracy (ASL)	.254	.103

** Correlation is significant at the 0.01 level (2-tailed).

Appendix I Code book

Table Appendix 12 Code book for justifications that rating a criterion “easy” or “very easy”

Code and subcodes	Memo	# Seg	# Docs
1 Identifying decision feature	Comments focusing on identifying a decision feature in the performance	0	0
1.1 Fluency	Comments related specifically to fluency	38	37
1.2 Quality of language	Comments related specifically to language quality	18	17
1.2.1 Mistakes	Highlighting ease of identifying mistakes	14	14
1.3 Quality of content	Comments related specifically to quality of content	22	21
1.4 Improves with practice	Comments related specifically to how ability to identify features improves with practice	4	3
1.5 General		4	3
2 Forming expectations of decision feature	Comments on how expectations based on experience and comparison of performances is helpful	14	9
3 Documenting decision feature	Comments underlining that the decision features can be documented quickly or easily in order to take a rating decision. e.g.: ticking off whether bullet points were addressed; writing down expression used	11	11
4 Decision feature is clear cut (correct/incorrect)	Comments highlighting the clear-cut nature of a decision feature. It is easy to decide whether something is right/wrong - correct/incorrect – there/not there	6	5
5 Descriptors are useful	Comments highlighting how the scale or descriptors are useful for rating	5	4
6 Fewer descriptors to deal with	Comments highlighting that the number of descriptors is easier to handle	3	3
7 Rater knowledge	Comments highlighting certain knowledge raters had they found helpful	3	2
8 Unspecified		10	9

Table Appendix 13 Code book for justifications that rating a criterion was difficult

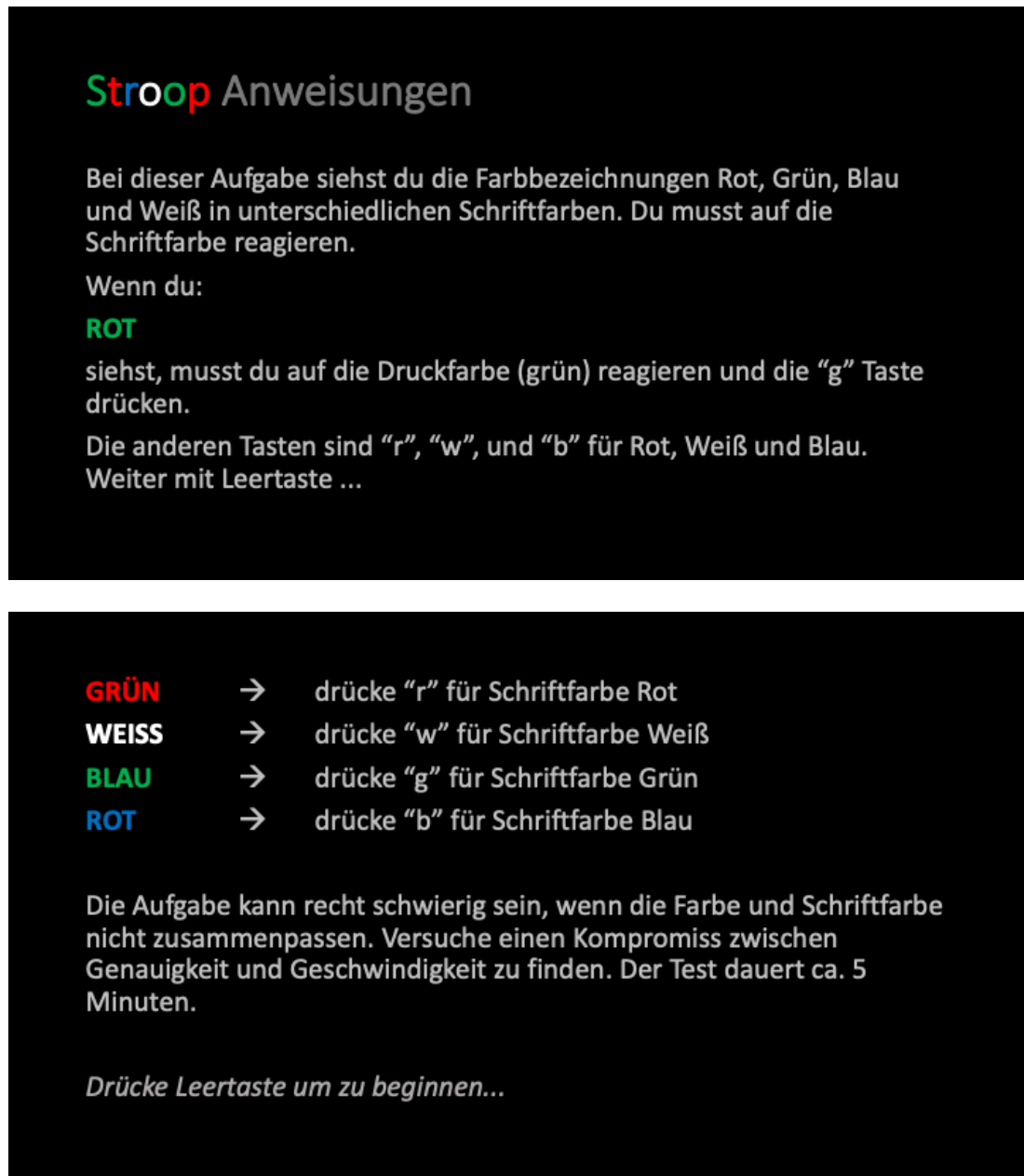
Code and subcodes	Memo	# Seg	# Docs
1 Difficulty with decision-making (no confidence in rating) Comments highlighting that raters are unsure about their decisions. The comments do not indicate issues with identifying or perceiving the decision feature(s).			
1.1 Performance is difficult to rate			
1.1.1 Uneven features	Raters comment that band as a whole does not match performance because: <ul style="list-style-type: none"> • features within the performance may be of varying quality • features within the performance varied over course of performance • features deemed important did not appear in performance 	44	33
1.1.2 Mismatch between scale and performance	Rating is problematic because: <ul style="list-style-type: none"> • performance was too short and interlocutor had to intervene • level of task achievement is unclear • some descriptor don't seem to apply • speakers stayed within their comfort zone 	14	14
1.2 Decision-making is flawed			
1.2.1 Subjectivity - harshness	Comments highlight that <ul style="list-style-type: none"> • it is difficult to rate certain features objectively • raters were unhappy about comparing performances 	16	15
1.2.2 Struggle with differentiating criteria	Comments highlighting how raters struggled to keep one criterion separate from other criteria.	25	20
1.3 Lacking rater knowledge (right/wrong, level, topic)			
1.3.1 What is B2?	Comments about <ul style="list-style-type: none"> what is sufficient for a level? what is sufficient for the B2 level? which words/structures should someone at B2 know? 	35	28
1.3.2 Interpretation of descriptors	Comments like <ul style="list-style-type: none"> descriptors sounded similar what is the difference between band 6 and 8? what does "quote from scale" mean? 	35	24

<p>2 Cognitive load and attention (no confidence in perception)</p> <p>Raters are unsure about their decisions and either mention complexity of criterion, noticing decision features and failing to hear features because they focussed on something else.</p>			
2.1 Identifying salient features or differences	<p>Commenting on not hearing/noticing certain features.</p> <p>as soon as raters say something like: I am not sure that what XY used/said was ENOUGH then it is more a question of their knowledge of the scale or level (→ 1.3.1) than noticing.</p>	21	17
2.2 Multi-tasking	Comments highlighting that raters struggled to hear features because they were paying attention to something else	21	18
2.3 Complexity of criterion	Raters mention that they struggle with rating the criterion as it has many descriptors and covers a broad construct.	8	6

Appendix J Cognitive tests

Appendix J.1 Stroop Test

Figure Appendix 4 Screenshots of Stroop instructions



Appendix J.2 Letters-Numbers Task

Figure Appendix 5 Screenshots of instructions for Letters-Numbers task

What to do in the following task?

In the following task you respond with button presses to letters and number. You will only need two keys (B and N)

You will always see a letter/number combination, for example **G1**.

If the letter/number combination appears at the top of the screen, you need to respond to the letter.

If the letter/number combination appears at the bottom of the screen, you need to respond to the number.

press space bar to continue

top

bottom

LETTER TASK

Consonant G,K,M,R ↓ press B	Vowel A,E,I,U ↓ press N
--------------------------------------	----------------------------------

NUMBER TASK

Odd 3,5,7,9 ↓ press B	Even 2,4,6,8 ↓ press N
--------------------------------	---------------------------------

If letter/number combination appears in top quadrants, respond to the letter (in this case, a "G").
If letter/number combination appears in bottom quadrants, respond to the number (in this case, a "6")

press space bar to continue

top

bottom

LETTER TASK

Consonant G,K,M,R ↓ press B	Vowel A,E,I,U ↓ press N
--------------------------------------	----------------------------------

NUMBER TASK

Odd 3,5,7,9 ↓ press B	Even 2,4,6,8 ↓ press N
--------------------------------	---------------------------------

So, in this case, you need to respond to the G and ignore the 6. The G is a consonant, so you press the **B** key!

press space bar to continue

top

bottom

	G4

LETTER TASK

Consonant G,K,M,R ↓ press B	Vowel A,E,I,U ↓ press N
--------------------------------------	----------------------------------

NUMBER TASK

Odd 3,5,7,9 ↓ press B	Even 2,4,6,8 ↓ press N
--------------------------------	---------------------------------

And in this case, you need to respond to the 4 (number) and ignore the G. The 4 is even, so you press the **N** key!

press space bar to continue

Now you should know everything you need to know for the most difficult part of this study.

Try to respond fast, and try to make few mistakes!

You should now be ready to go!

Press Q to start,
or use up and down arrows to go back
to previous pages...

Appendix J.3 Digit Span Task

Figure Appendix 6 Instructions for digit span task

ANLEITUNG

1. Du wirst jetzt eine Reihe von Zahlen hören.
2. Sobald die Reihe zu Ende ist, erscheint die Anweisung "Aufschreiben".
3. Notiere dann sofort die Zahlen in derselben Reihenfolge, wie du sie gehört hast.

Appendix J.4 Keep track

Figure Appendix 7 Instructions and sample item for Keep Track task

ANLEITUNG

1. Für diese Aufgabe werden dir Wörter aus sechs Kategorien gezeigt: Metalle, Farben, Tiere, Länder, Familie und Währung.
2. Zuerst erfährst du die Kategorien, für welche du dir Wörter merken musst.
3. Dann werden dir kurz hintereinander 15 Wörter gezeigt.
4. Abschließend musst du das jeweils zuletzt genannte Wort der Zielkategorien aufschreiben.

Metalle Aluminium Blei Eisen	Farben rot grün gelb	Tiere Kuh Pferd Katze
Länder Deutschland Italien Schweiz	Familie Mutter Vater Schwester	Währung Pfund Euro Franken

ÜBUNGSLAUF

Du siehst nun 15 Wörter in rascher Abfolge.

Merke dir das zuletzt genannte Wort der folgenden Kategorien:

- Metalle
- Familie
- Währung

Weiter mit Rechtsklick...

Italien

Metalle
Familie
Währung

Appendix J.5 Trail Making Test (Part A)

Trail Making Test A und B (Aufmerksamkeit, exekutive Funktionen)

Zahlen (Test A) oder Zahlen und Buchstaben (Test B) sollen in aufsteigender Reihenfolge so schnell wie möglich verbunden werden. Die benötigte Zeit wird gestoppt.

Übungsblatt A

„Auf diesem Blatt hat es verschiedene Nummern. Beginnen Sie bei der Zahl 1 und zeichnen Sie einen Strich von 1 nach 2, von 2 nach 3, von 3 nach 4 usw., bis Sie am Ende sind.“

Während der Erklärung Schritte zeigen, von 1 - Ende

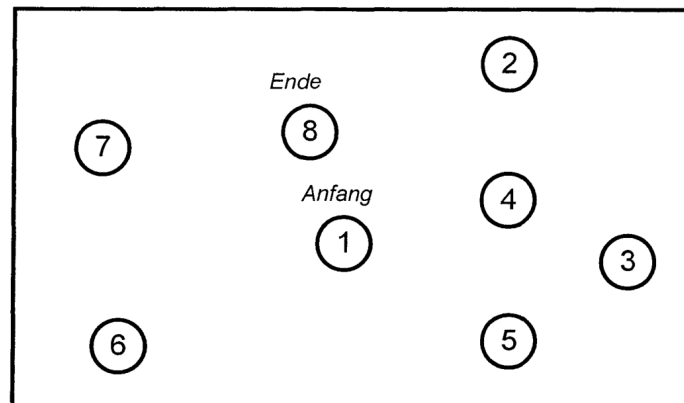
„Zeichnen Sie die Linien so schnell wie möglich und entfernen Sie den Bleistift nicht vom Papier!“

Wenn Übungsblatt A verstanden wurde, zur Testung übergehen.

Falls die TP im Übungsbeispiel einen Fehler macht, wird sie sofort darauf aufmerksam gemacht!

Trail Making Test A:

Übungsbeispiel



Zeit Test A Sek.

Fehler Test A

9. Trail Making Test A

„Bitte verbinden sie alle Zahlen von 1 bis 25 in aufsteigender Reihenfolge. Zeichnen Sie die Linien so schnell wie möglich ein und entfernen Sie den Bleistift nicht vom Papier!“

Erst jetzt das Blatt geben.

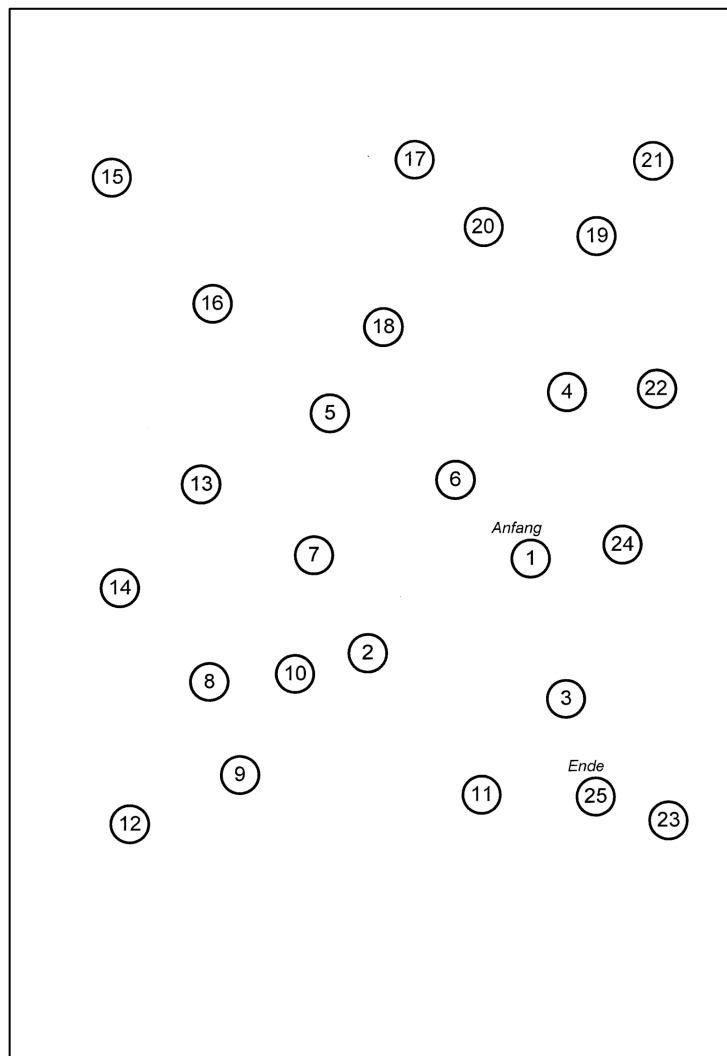
„Hier ist ihr Anfang. Los!“

Parallel zum „Los!“ Stoppuhr drücken.

Wenn die TP einen Fehler macht, sofort darauf aufmerksam machen und Fehler korrigieren lassen, d.h. zum letzten richtigen Kreis zurückkehren und von dort aus weiterfahren. Zeit weiter laufen lassen.

Bewertung: Zeit (in Sek.); Fehler werden vor allem indirekt, durch die zusätzlich aufgewendete Zeit bewertet. Die Striche der TP sollen die Kreise mindestens berühren (darauf hinweisen, zählt aber nicht als Fehler).

Abbruch Test A nach 3 Min.



Appendix J.6 Trail Making Task (Part B)

Übungsblatt B

„Zeichnen Sie bitte eine Linie von 1 nach A, von A nach 2, von 2 nach B, von B nach 3, von 3 nach C usw., Zahlen in aufsteigender Reihenfolge, Buchstaben nach dem Alphabet.“

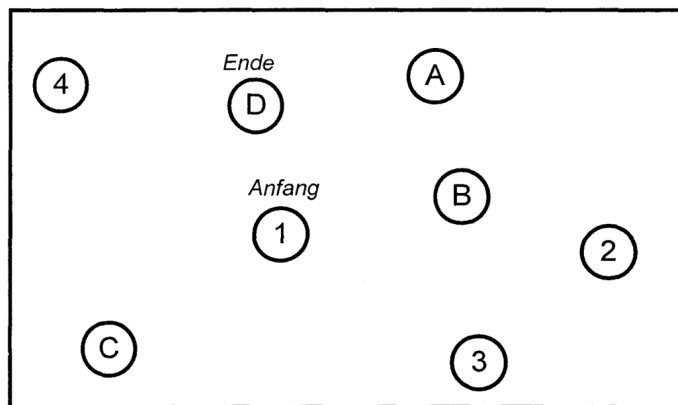
Während der Erklärung jeden Schritt von 1 - A bis Ende zeigen.

„Zeichnen Sie die Linien so schnell wie möglich ein, ohne den Bleistift vom Blatt zu nehmen.“

Wenn Übungsblatt B verstanden wurde zur Testung übergehen.

Trail Making Test B

Übungsbeispiel



Zeit Test B Sek.

Fehler Test B

10. Trail Making Test B

„Erinnern Sie sich daran, dass Sie zuerst eine Zahl, dann einen Buchstaben, dann wieder eine Zahl, dann wieder einen Buchstaben, usw. verbinden müssen. Halten Sie die Reihenfolge ein und lassen Sie keine Kreise aus. Zeichnen Sie die Linien so schnell wie möglich ein, ohne den Bleistift vom Blatt zu nehmen.“

Unbedingt darauf hinweisen, dass die Buchstabenreihenfolge i j k lautet.
Erst jetzt Blatt geben.

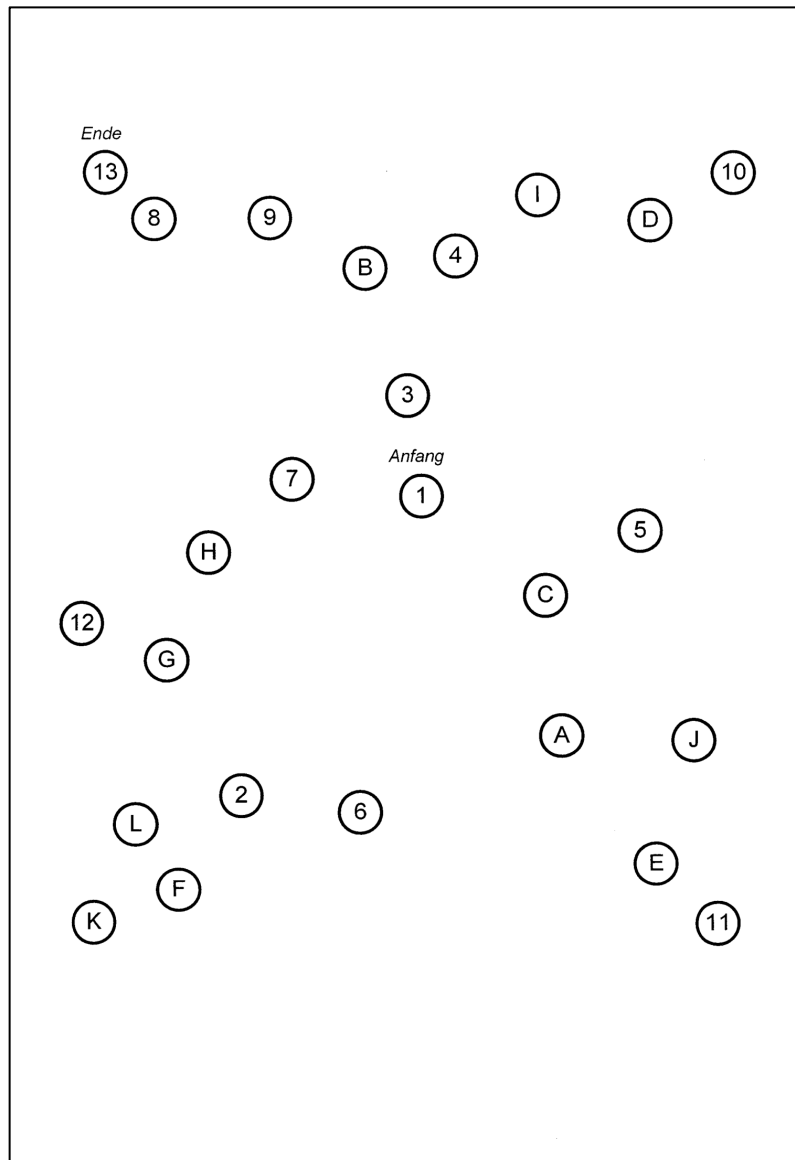
„Hier ist ihr Anfang. Los!“

Parallel zum „Los!“ Stoppuhr drücken.

Wenn die TP einen Fehler macht, sofort darauf aufmerksam machen und Fehler korrigieren lassen, d.h. zum letzten richtigen Kreis zurückkehren und von dort aus weiterfahren. Zeit weiter laufen lassen.

Bewertung: Zeit (in Sek.); Fehler werden vor allem indirekt, durch die zusätzlich aufgewendete Zeit bewertet. Die Striche der TP sollen die Kreise mindestens berühren (darauf hinweisen, zählt aber nicht als Fehler).

Abbruch Test B nach 5 Min.



Appendix K General Decision-Making Style Inventory (Scott & Bruce, 1995)

Appendix K.1 GDMSI Items

GDMS INVENTORY

BELOW ARE STATEMENTS DESCRIBING HOW PEOPLE GO ABOUT MAKING IMPORTANT DECISIONS. PLEASE INDICATE HOW MUCH YOU AGREE OR DISAGREE WITH EACH STATEMENT

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1. I double-check my information sources to be sure I have the right facts before making decisions. (R)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. When making decisions, I rely on my instincts. (I)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. I often need the assistance of other people when making important decisions. (D)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. I avoid making important decisions until the pressure is on. (A)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. I generally make snap decisions. (S)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. I make decisions in a logical and systematic way. (R)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. When I make decisions, I tend to rely on my intuition. (I)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. I rarely make important decisions without consulting other people. (D)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. I postpone decision making whenever possible. (A)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. I often make decisions on the spur of the moment. (S)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11. My decision making requires careful thought. (R)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12. I generally make decisions that feel right to me. (I)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13. If I have the support of others, it is easier for me to make important decisions. (D)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14. I often procrastinate when making important decisions. (A)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15. I make quick decisions. (S)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16. When making a decision, I consider various options in terms of a specific goal. (R)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17. When I make a decision, it is more important for me to feel the decision is right than to have a rational reason for it. (I)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
18. I use the advice of other people in making my important decisions. (D)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
19. I generally make important decisions at the last minute. (A)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
20. I often make impulsive decisions. (S)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
21. I explore all of my options before making a decision. (R)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
22. When I make a decision, I trust my inner feelings and reactions. (I)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
23. I like to have someone steer me in the right direction when I am making important decisions. (D)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
24. I put off making many decisions because thinking about them makes me uneasy. (A)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
25. When making decisions, I do what seems natural at the moment. (S)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Note. There were no DMS labels in the questionnaires for the participants. R = Rational, I = Intuitive, D = Dependent, A = Avoidant, S = Spontaneous. From "Decision- Making Style: The Development and Assessment of a New Measure," by S. G. Scott & R. A. Bruce, 1995, *Educational and Psychological Measurement*, 55, p. 826. Sequence of items and presentation from "Individual differences in rater decision-making style: An exploratory study," by B. Baker, 2012, *Language Assessment Quarterly*, 9, p. 248.

Appendix K.2 GDMSI Coding

Scott & Bruce's GDMSI (1995) includes five items per decision-making style. No items need to be reversed as they all align in terms of direction. The original publication does not provide a recommended sequence of items. The layout for this experiment was based on Baker's (2012) study.

Table Appendix 14 Map of items per DMS subscale

DMS	Items
Rational	1, 6, 11, 16, 21
Intuitive	2, 7, 12, 17, 22
Dependent	3, 8, 13, 18, 23
Avoidant	4, 9, 14, 19, 24
Spontaneous	5, 10, 15, 20, 25

Appendix L Rational-Experiential Inventory (Pacini & Epstein, 1999)

Appendix L.1 REI-40 Items

Rational Experiential Inventory (REI-40)

Participant: _____

Please use the following scale (1-5) to answer the questions.	completely false		completely true		
	(1)	(2)	(3)	(4)	(5)
1. I have a logical mind. (RA)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. I prefer complex problems to simple problems. (RE)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. I believe in trusting my hunches. (EA) <small>(hunch = Gefühl, Vermutung)</small>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. I am not a very analytical thinker. (RA-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. I trust my initial feelings about people. (EA)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. I try to avoid situations that require thinking in depth about something. (RE-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. I like to rely on my intuitive impressions. (EE)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. I don't reason well under pressure. (RA-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. I don't like situations in which I have to rely on intuition. (EE-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. Thinking hard and for a long time about something gives me little satisfaction. (RE-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11. Intuition can be a very useful way to solve problems. (EE)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12. I would not want to depend on anyone who described himself or herself as intuitive. (EE-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13. I am much better at figuring things out logically than most people. (RA)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14. I usually have clear, explainable reasons for my decisions. (RA)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15. I don't think it is a good idea to rely on one's intuition for important decisions. (EE-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16. Thinking is not my idea of an enjoyable activity. (RE-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17. I have no problem thinking things through carefully. (RA)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
18. When it comes to trusting people, I can usually rely on my gut feelings. (EA)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
19. I can usually feel when a person is right or wrong, even if I can't explain how I know. (EA)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
20. Learning new ways to think would be very appealing to me. (RE)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please use the following scale (1-5) to answer the questions.	completely false			completely true	
	(1)	(2)	(3)	(4)	(5)
21. I hardly ever go wrong when I listen to my deepest gut feelings to find an answer. (EA)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
22. I think it is foolish to make important decisions based on feelings. (EE-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
23. I tend to use my heart as a guide for my actions. (EE)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
24. I often go by my instincts when deciding on a course of action. (EE)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
25. I'm not that good at figuring out complicated problems. (RA-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
26. I enjoy intellectual challenges. (RE)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
27. Reasoning things out carefully is not one of my strong points. (RA)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
28. I enjoy thinking in abstract terms. (RE)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
29. I generally don't depend on my feelings to help me make decisions. (EE-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
30. Using logic usually works well for me in figuring out problems in my life. (RA)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
31. I think there are times when one should rely on one's intuition. (EE)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
32. I don't like to have to do a lot of thinking. (RE-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
33. Knowing the answer without having to understand the reasoning behind it is good enough for me. (RE-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
34. Using my gut feelings usually works well for me in figuring out problems in my life. (EA)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
35. I don't have a very good sense of intuition. (EA-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
36. If I were to rely on my gut feelings, I would often make mistakes. (EA-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
37. I suspect my hunches are inaccurate as often as they are accurate. (hunch = Gefühl, Vermutung) (EA-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
38. My snap judgements are probably not as good as most people's. (EA-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
39. I am not very good at solving problems that require careful logical analysis. (RA-)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
40. I enjoy solving problems that require hard thinking. (RE)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Note.</i> Labels were not in questionnaires for participants. RA = Rational Ability, RE = Rational Engagement, EA = Experiential Ability, EE = Experiential engagement. Labels including a minus sign (-) need to be reverse coded. Subscale scores are based on averaging responses to 10 items of subscale.					

Appendix L.2 REI-40 Coding

Items to reverse: 4, 6, 8, 9, 10, 12, 15, 16, 22, 25, 27, 29, 32, 33, 35, 36, 37, 38, 39

Table Appendix 15 Item map of REI-40 and subscales

REI- Subscale	Items
Rational Ability	1, 4, 8, 13, 14, 17, 25, 27, 30, 39
Rational Engagement	2, 6, 10, 16, 20, 26, 28, 32, 33, 40
Experiential Ability	3, 5, 18, 19, 21, 34, 35, 36, 37, 38
Experiential Engagement	7, 9, 11, 12, 15, 22, 23, 24, 29, 31

Appendix M Participant Information



Informationsblatt

Title: Exploring decision-making in rating speaking

Researcher: Kathrin Eberharter (kathrin.eberharter@uibk.ac.at)

Ich arbeite als Forschungsassistentin an der Universität Innsbruck und bin zugleich PhD Studentin an der Lancaster University. Sie wurden zur Teilnahme an einem Forschungsprojekt eingeladen. Bitte lesen Sie die folgenden Informationen aufmerksam durch, bevor Sie Ihre Teilnahme per Einverständniserklärung bestätigen.

Was ist der Zweck dieser Studie?

Die Studie wird im Rahmen meines Doktoratsstudiums am *Department of Linguistics and English Language* der *Lancaster University* durchgeführt. Das Ziel der Studie ist es zu untersuchen, inwiefern individuelle Entscheidungspräferenzen das Bewertungsverhalten bei mündlichen Englischprüfungen beeinflusst.

Was umfasst eine Teilnahme bei dieser Studie?

Bei der Studie bewerten Englisch LA Studierende mündliche Performanzen anhand des österreichischen Bewertungsrasters. Eine kleinere Gruppe von TeilnehmerInnen nimmt auch an einer zweiten kleineren Studie teil, wo mittels Eye-Tracking Technologie die Interaktion von PrüferIn und Bewertungsraster während einer Performanz aufgezeichnet wird.

Warum wurde ich zur Teilnahme eingeladen?

Sie wurden angesprochen, weil Sie sich am Ende Ihres Studiums befinden und einen ähnlichen Informationsstand wie viele österreichische Lehrpersonen haben.

Was passiert, wenn ich einer Teilnahme zustimme?

Eine Teilnahme umfasst dies die folgenden Punkte: Sie bewerten 30 Performanzen (jeweils ca. 5 Minuten) mit Hilfe eines Bewertungsrasters und einer speziellen Software an einem Computer der Universität Innsbruck. Des weiteren füllen Sie zwei kurze Fragebögen aus und nehmen an einem kurzen Interview bestehend aus fünf standardisierten Arbeitsgedächtnisaufgaben teil (ca. 35 Minuten). Eine kleine Auswahl an Personen wird des Weiteren zu einer Bewertungsrunde von fünf Performanzen vor einem Eye-Tracker eingeladen.

Was sind die Vorteile einer Teilnahme?

Sie erhalten ohne großen Aufwand ein grundlegendes Training in der Anwendung des Bewertungsrasters und können viel Praxis bei der Anwendung des Bewertungsrasters auf unterschiedlichste Sprechperformanzen sammeln. Sie erhalten auch individuelles Feedback zur Qualität Ihrer Bewertungen und inwiefern sich diese mit der Einschätzung Ihrer KollegInnen und einem Expertenteam überschneiden. Darüberhinaus trägt Ihre Teilnahme zu einer ersten Annäherung an eine große Forschungslücke bei und hilft näher zu erschließen, wie unterschiedliche Individuen eine Bewertungsskala verwenden und ihre Bewertungsentscheidungen treffen.

Was sind die möglichen Nachteile und Risiken einer Teilnahme?

Die Teilnahme an der ersten Studie ist mit ca. 6 Stunden relativ aufwendig. TeilnehmerInnen der Folgestudie müssen mit weiteren 2 Stunden Zeitaufwand rechnen.

Was passiert, wenn Sie nicht teilnehmen wollen, oder Ihre Teilnahme abbrechen?

Sie können innerhalb eines Monats nach der Datensammlung ohne Angabe von Gründen zurücktreten. Da gewisse statistische Modelle bei der Auswertung der Ergebnisse angewendet werden, kann es sein, dass der Datensatz mit Ihren Informationen zu einem späteren Zeitpunkt nur mehr teilweise aus der Studie entfernt werden kann.

Ist meine Teilnahme vertraulich?

Alle Informationen die während dieser Studie gesammelt werden, werden strikt unter Verschluss gehalten. Bei jeglicher Veröffentlichung der Ergebnisse (Gespräche mit anderen Forschern, Dissertation, Artikel, Präsentationen, Buchkapitel, etc.) wird Ihr Name durch ein Pseudonym ersetzt oder gänzlich ausgespart. Alle digitalen Daten werden auf Passwort-geschützten Geräten und Festplatten abgelegt. Sämtliche gedruckte Dokumente werden in einem verschlossenen Schrank in meinem Büro an der Universität aufbewahrt. Gemäß der Vorgaben der Lancaster University werden alle Daten für die Dauer von 10 Jahren archiviert.

Was passiert mit den Ergebnissen der Studie?

Die Ergebnisse werden nur für wissenschaftliche und akademische Zwecke verwendet. Dies umfasst meine Dissertation und weitere Publikationen, wie z.B. Artikel in Fachzeitschriften, als auch Präsentationen bei nationalen und internationalen Kongressen.

Was tun bei Fragen?

Falls Sie irgendwelche Fragen haben oder unzufrieden sind mit Ihrer Teilnahme und dem Verlauf der Studie, kontaktieren Sie bitte mich oder auch die Leiterin des Instituts, Professor Elena Semino in Kontakt (Department of Linguistics and English Language, e.semino@lancaster.ac.uk, +44 (0)1524 594176).

Diese Studie wurde von Mitgliedern des Lancaster University Ethics Committee (UREC) begutachtet und zugelassen.

Danke, dass Sie eine Teilnahme in Erwägung ziehen.

Weitere Informationen und Kontakt

Kathrin Eberharter (kathrin.eberharter@uibk.ac.at)
Wissenschaftliche Mitarbeiterin
Universität Innsbruck
Institut für Fachdidaktik/Abteilung Didaktik der Sprachen
Innrain 52d, 6020 Innsbruck
Telefon: +43 512 507-4305

Appendix N Consent Form



Einverständniserklärung

Projekttitel: Exploring decision-making in rating speaking

Ich habe das Informationsblatt bezüglich der Studie gelesen und verstanden. Offene Fragen bezüglich des Projekts wurden von Frau Eberharter zu meiner Zufriedenheit beantwortet. ☐

Der Zweck des Projekts und die Anforderung an mich als TeilnehmerIn wurden mir erklärt. Ich stimme den Aspekten des Informationsblattes zu, die meine Teilnahme beschreiben. ☐

Ich stimme zu, dass meine Bewertungen aufgezeichnet werden. Für den Fall, dass ich zur Folgestudie eingeladen werde, stimme ich zu, dass das Interview aufgezeichnet wird. ☐

Mir ist bewusst, dass meine Teilnahme freiwillig ist und dass ich bis zu einem Monat nach Ende der Datensammlung das Recht habe, vom Projekt zurückzutreten. Ein späterer Austritt aus der Studie ist eventuell nicht mehr möglich, wenn alle Daten für die Analyse bereits aggregiert und weiter verarbeitet wurden. ☐

Mir wurde erklärt, dass alle Daten mit einem Pseudonym versehen werden und meine Identität an keinem Punkt an Dritte offengelegt wird. ☐

Mir wurde eine Kopie dieser Einverständniserklärung und des Informationsblattes übermittelt. ☐

Ich stimme der Teilnahme an dieser Studie zu. ☐

Name:

Unterschrift:

Datum:

Appendix O Summary of Study 2 Results

Table Appendix 16 Summary of findings for cognitive tasks, REI-40 and GDMSI with rater severity

		Rater severity				
		Full model	TA	FLIN	RSL	ASL
Cognitive tasks	Stroop	-.20	-.18	-.20	-.22	-.20
	Letter-numbers	-.01	.06	.01	-.15	-.06
	Digit Span	-.05	-.03	-.04	-.02	-.06
	Keep Track	-.17	-.16	-.13	-.20	-.15
	Trail Making	-.15	-.01	-.08	-.25	-.21
Processing mode (REI-40)	Rationality	-.04	.15	-.07	-.13	-.06
	Experientiality	-.09	-.08	-.01	-.15	-.09
Decision-making style (GDMSI)	Rational	.28	.24	.33	.16	.21
	Intuitive	-.39	-.24	-.32	-.48**	-.40
	Dependent	-.38	-.36	-.27	-.39	-.33
	Avoidant	-.52**	-.38	-.41	-.51**	-.50**
	Spontaneous	-.16	-.03	-.11	-.15	-.19
Rating behaviour	Deliberation time	-.18	-.08	-.28	-.19	-.09
	Time to first decision	-.18	-.25	-.07	-.07	-.25
	Revisions	-.01	.07	-.13	-.07	.10

Note. TA = Task achievement, FLIN = Fluency, RSL = Range of spoken language, ASL = Accuracy of spoken language. p-values corrected via Holm-Bonferroni stepwise procedure.

Table Appendix 17 Summary of results for cognitive tasks, REI-40 and GDMSI with rater accuracy

		Rater accuracy				
		Full model	TA	FLIN	RSL	ASL
Cognitive tasks	Stroop	.23	.14	.27	.23	.15
	Letter-numbers	.06	-.13	.18	.17	.07
	Digit Span	-.01	-.11	-.02	-.03	.07
	Keep Track	.03	-.17	.04	.11	.21
	Trail Making	.06	-.05	-.04	.19	.22
Processing mode (REI-40)	Rationality	.05	-.13	.17	.17	.01
	Experientiality	.04	-.08	-.10	.09	.15
Decision-making style (GDMSI)	Rational	-.14	-.09	-.09	-.14	-.14
	Intuitive	.29	.02	.16	.42	.39
	Dependent	.31	.13	.15	.39	.29
	Avoidant	.37	.09	.25	.53**	.31
	Spontaneous	.00	-.14	.04	.10	.03
Rating behaviour metrics	Deliberation time	.12	-.01	.21	.12	.10
	Time to first decision	-.17	-.44	.03	-.04	-.05
	Revisions	.39	.56**	.22	.17	.23

Note. TA = Task achievement, FLIN = Fluency, RSL = Range of spoken language, ASL = Accuracy of spoken language. p-values corrected via Holm-Bonferroni stepwise procedure.

Table Appendix 18 Summary of results for cognitive tasks, REI-40 and GDMSI with rater behaviour metrics

		Rater behaviour metrics		
		Deliberation time	Time to first decision	Revisions
Cognitive tasks	Stroop	.37	.18	.13
	Letter-numbers	-.25	-.19	-.05
	Digit Span	.02	-.06	-.01
	Keep Track	.11	.21	-.37
	Trail Making	-.29	-.18	-.10
Processing mode (REI-40)	Rationality	.04	.12	-.07
	Experientiality	-.20	.03	-.08
Decision-making style (GDMSI)	Rational	.37	.27	.05
	Intuitive	-.07	-.02	-.06
	Dependent	.15	-.15	.29
	Avoidant	.03	-.13	.25
	Spontaneous	-.36	-.23	.05

Note. p-values corrected via Holm-Bonferroni stepwise procedure.