DE GRUYTER OPEN

ICAME Journal, Volume 38, 2014, DOI: 10.2478/icame-2014-0001

Assessing the inter-coder reliability of the Body Type Dictionary (BTD)

Laura A. Cariola, Lancaster University

Abstract

Computer-assisted content analysis has many advantages compared to a manual scoring system, provided that computerised dictionaries represent valid and reliable measures. This study aimed to assess the inter-coder reliability, alternate-form reliability and scoring consistency of the Body Type Dictionary (BTD) (Wilson 2006) based on Fisher and Cleveland's (1956, 1958) manual body boundary scoring scheme. The results indicated an acceptable inter-coder agreement with barrier and penetration imagery in the sub-sample (N = 53) of manually coded Rorschach responses. Additionally manually coded scores showed an acceptable correlation with the computerised frequency counts, and thus indicating an alternate-form reliability. In the full data set (N = 526), barrier imagery in the Rorschach responses only correlated with the picture response test, showing low scoring consistency, which might disconfirm the notion of body boundary awareness representing a stable personality trait but instead it might be dependent on the level of cognitive dedifferentiation.

1 Introduction

Recent developments in psychological and linguistic research show an increased interest in exploring how perceptions of personal experiences are related to individuals' body awareness (e.g. Wilson 2009). Fisher and Cleveland (1956, 1958) devised the Body Image scoring system to examine how individual differences in body boundary awareness relate to verbal expressions of experiential perceptions. The Body Type Dictionary (BTD) (Wilson 2006) represents a computerised version of Fisher and Cleveland's Body Image manual scoring system, which has been applied to a variety of text types, such as fantasy stories and religious texts. This study aims to assess the inter-coder reliability, alternate-form reliability and scoring consistency of the BTD as a means to ensure the robustness and reproducibility of its lexical content.

1.1 Inter-coder reliability

The construction of a content analysis coding scheme relies primarily on the researcher's judgment regarding how to code the lexical content of a coding category. The quantitative assessment of the reliability of a coding scheme verifies that "the obtained ratings are not idiosyncratic results of the coders' subjective judgment" (Tinsley and Weiss 1975: 359). Although reliability has been widely neglected in content analysis studies (Krippendorff 2004), reliability assessments are important. The lexical classification of a content analysis coding scheme that is not sufficiently reliable might produce results that are not regarded as valid, which would yield meaningless data interpretations (Weber 1990; Singletary 1994; Potter and Levine-Donnerstein 1999; Lombard, Snyder-Duch and Bracken 2010). Thus, the application of content analysis has often been criticised due to its lack of reliability with regard to ensuring an acceptable scientific standard (Neundorf and Skalski 2010).

A content analysis coding scheme is deemed reliable to the extent that different coders have a shared understanding of the lexical content and classification categories that result in a high coding agreement (Neuendorf 2002). This high coding agreement indicates that the lexical content of the coding scheme is accurate and consistent with the underlying theoretical construct it aims to measure. A low inter-coder agreement, on the other hand, might be indicative of ambiguities and weaknesses in the lexical content, as well as inaccuracies related to insufficient training of the coders, cognitive differences among the coders, ambiguities in the coding instructions, or weaknesses in the research methodology based on an insufficient theoretical foundation (Weber 1990; Kolbe and Burnett 1991).

A sufficient inter-reliability agreement based on manual annotation of the lexical coding scheme would be indicative of the reliable application of a computerised measurement that is theoretically based on the same lexical scoring scheme. Such a computerized scoring would then represent a parallel coding scheme that would result in a high strength of associations with the manual scoring when applied to the same texts and thus indicating an alternate-form reliability (Jackson 2011). Repeated coding of the same text using the same reliable and valid lexical coding content scheme would then result in consistently replicable results (Weber 1990; Rourke *et al*. 2000), for which computerized coding provides the advantage of reliably producing the same frequency of lexical content in a time efficient manner.

1.2 The body image concept

Fisher and Cleveland's (1956, 1958) manual body boundary scoring system represents a lexical measurement that assesses the unconscious process, revealing, "the degree of definiteness the individual assigns to his body boundaries" (1958: 57). Variations in body boundary awareness have been investigated across a wide range of psychological phenomena, including body self-schema, psychosomatic illnesses, achievement motivation, stress and coping, and psychopathology (for a detailed overview, see Fisher 1986). Fisher and Cleveland (1956, 1958) firstly proposed a valid and reliable content-analysis scoring system of body boundary awareness based on verbal responses of Rorschach inkblot tests (for a detailed summary regarding the reliability and validity of the body boundary scoring system see O'Neill 2005). A detailed assessment of the surface and boundary descriptions for these inkblot responses revealed that responses could be differentiated into two scoring categories, which are 'barrier imagery' and 'penetration imagery'. Barrier imagery responses emphasised the positive features of definite structure, substance and surface qualities of the boundary peripheries of objects. The definite boundary qualities that are reflected in barrier responses describe the protective, enclosing, decorative, or concealing qualities of a surface. In contrast, penetration imagery responses reflect a lack of these protective and enclosing boundaries by emphasising the sensation of fragility, permeability, openness and destruction of definite boundaries. According to this scoring system, a high frequency of boundary imagery corresponds to a High Barrier personality, whereas a low frequency of barrier imagery indicates a Low Barrier personality. However, both personality categories are assumed to represent related personality dimensions, rather than opposite ends of a polar personality model.

The body image scoring system has been used in qualitative and quantitative studies to investigate body boundary distortion in pathological and non-pathological forms of altered states of consciousness (ASC). Weak body boundaries in patients diagnosed with schizophrenia can be reflected in their psychotic delusions, which may include transgressions and vagueness with regard to their body boundaries, such as feelings of depersonalisation and changes in body consistency (Guimon 1997). The blurring of body boundaries in psychotic disorders represents a phenomenological characteristic that is also associated with non-pathological forms of ASC. For example, with regard to extra-sensory perceptions (ESP), individuals who have high scores on ESP showed lower body boundary definiteness (i.e. higher penetration and lower barrier imagery scores) than individuals who have low ESP scores (Schmeidler and LeShan 1970). Similarly, body boundary definiteness was lower in hypnotised individuals than in

individuals experiencing ordinary states of consciousness (Saraceni, Ruggeri and Filocamo 1980). Such changes of body boundary awareness have been associated with levels of regressive cognitive functioning. For example, Buck and Barden (1971) found that the frequencies of penetration imagery would increase in the expected direction of conceptual to primordial thought functioning - autobiographical report, daydreams, and dreams. Such a relationship between penetration imagery and primordial thought language has also been identified in relation to religious-mystical experiences (Wilson 2009; Cariola 2012). Theoretical models similar to Fisher and Cleveland's High and Low Barrier personality categories have been proposed, including skin ego (Anzieu 1985), amoebic self-theory (Burris and Rempel 2004), secondary skin formation (Bick 1964; Ogden 1989), and crustacean and amoebid self-protection in infants with autism (Tustin 1981), among other theories.

1.3 Body Type Dictionary (BTD)

The BTD (Wilson 2006) is a computerised dictionary that calculates the frequency of semantic items that are categorised as barrier imagery and penetration imagery based on Fisher and Cleveland's (1956, 1958) manual scoring system of High and Low Barrier personalities. The BTD contains 551 barrier imagery words, 231 penetration imagery words, and 70 exception words, which prevent the erroneous matching of ambiguous word stems assigned to 12 semantic categories (Wilson 2009) (cf. Appendix 1). Whereas Fisher and Cleveland's manual scoring system equated the frequencies of individual lexical items and context-dependent phrases, the computerised coding of the BTD's barrier and penetration imagery lexis is context-independent. Due to these inherent technical differences between the computerised and manual scoring schemes, the lexical content of the BTD represents a more restricted scope of semantic categories and lexical items as compared to Fisher and Cleveland's manual scoring system. For example, the BTD excludes polysemous words (e.g. *well*) and shelled sea animals due to their relation with seafood dishes (e.g. Lobster Thermidor). In the latter example, the use of barrier and penetration imagery is then related to convention, such as the name of a culinary dish, whereas, in the former example, in particular, the BTD scores individual words that are assumed to represent either the barrier or penetration imagery or adverb, or an adverb that would not be categorized with the body boundary imagery classification. However, the BTD's tagging capacity is limited in that it is not able to identify and classify barrier and penetration related meanings in phrase-based lexical content, whereas Fisher and Cleveland's manual scoring system is able to do so. For example, the manual dictionary would classify the expression, 'squirrel run over' as penetra-

tion imagery, whereas the computerised coding would not code any of the lexical items in this expression as penetration imagery, nor would it be able to decode the denoted mental image of the described destruction of the animal.

1.4 Hypotheses

Although the Body Type Dictionary (BTD) has been used in a variety of studies, it has not been assessed with regard to whether a) its semantic lexicon accurately measures barrier and penetration imagery, and b) repeated measures taken under the same conditions would reflect reproducible results with regard to the barrier and penetration imagery frequencies. The first part of this study aimed to assess the inter-coder reliability of the BTD by applying a manual coding of the body boundary imagery. The second part of this study explored the alternate-form reliability of the BTD by comparing manual and computerized coding. The third part aimed to assess the scoring consistency by measuring the association between computerized coded barrier, penetration and sum body boundary imagery across all of the experimental conditions, i.e. responses to the Rorschach and picture response test, the narratives of everyday memories and dream memories, and dream interpretations. Thus, for the first experiment of this study, it was predicted that (H1) manually coded barrier, penetration and sum body boundary imagery would demonstrate an acceptable inter-coder agreement. The second hypothesis (H2) of this study was based on the prediction that manual measures of barrier, penetration and sum body boundary imagery would be significantly and positively correlated with the computer-assisted measures of the same linguistic variable (i.e. manual measures of barrier imagery with the computer-assisted measures of barrier imagery, etc.), and thus indicating alternate-form reliability. The assessment of consistency of computerized scoring (H3) was based on the assumption that computer-assisted frequency measures for the linguistic variables (i.e. barrier, penetration and sum body boundary imagery) would be significantly correlated with the frequency measures for the same linguistic variables across all of the experimental conditions.

2 Method

2.1 Participants

The participants in this study were recruited from an e-mail that was sent to a number of academic departments within the majority of British Universities and subsequently the e-mail was distributed to the students. A total of 769 native British English speakers participated in the study, although 243 participants who provided incomplete or irrelevant responses were removed from the sample. In

total, the responses of 526 participants (358 females, l68 males) aged between l7–64 years (M = 25.47, SD = 10.63)[2] were used for further analysis, of which 526 participants provided responses to the Rorschach and picture response task, 488 participants provided a written narrative regarding an everyday memory, 450 participants provided a written narrative regarding a dream memory, and 427 participants provided an interpretation of a recalled dream memory.

2.2 Experimental procedure

The online survey was produced with the web-based software Survey Monkey (http://www.surveymonkey.net). The study's online questionnaire included an initial briefing that outlined the purpose of the research project. Once participants decided to participate in the experiment, they disclosed their demographic information, including gender, age, and native language. Then, participants were asked to write open-ended written responses to three types of experimental conditions, as follows: two types of projective tests (i.e. Rorschach inkblot test and picture response task), two types of memory recall tasks (i.e. an everyday memory recall and a dream recall), and a dream interpretation task. Completion of the experiment was not timed, and participants were informed that they could re-enter and complete their survey at any time. At the end of the experiment, participants were thanked and presented with a debriefing that explained the purpose of the study. The study obtained full ethical approval by the Ethics Committee at Lancaster University.

2.3 Stimuli

The following two different types of projective tests were used in this study: the Rorschach inkblot test (Rorschach 1921) and a picture response test (as an alternative to the TAT test). The Rorschach inkblot test represents a traditional projective test based on the presentation of ten symmetrically shaped inkblots, of which seven inkblots are black-and-white and the remaining three inkblots are in colour. The picture response test used in this study was based on four photographs. In this experiment, participants were presented with both the Rorschach inkblot test and the picture response test on a computer screen and then asked to write down a short interpretation of the inkblot and pictures in open-ended answer comment boxes. Whereas the Rorschach test is based on the analysis of participants' freely-associated interpretations of the inkblot percepts, the original TAT test (Morgan and Murray 1935) typically presents a set of drawings that participants are asked to freely associate with a narrative that follows a classical Aristotelian narrative structure (i.e. definite beginning, middle and ending). For the purpose of this study, four pictures were selected that were related to the

implied visual ambiguity of barrier and penetration imagery (see Figures 1–4 in Appendix 2). The pictures were selected according to their visual body boundary content, which included barrier imagery (e. g. clothing items) and penetration imagery (e.g. bombarded houses). The pictures are aimed to elicit freely associated narratives that would provide insight into Fisher and Cleveland's (1956, 1958) assumption that individuals project their own body boundary awareness onto external perceptions. Based on this assumption, the narratives of High Barrier personality types would reflect an inflated body boundary imagery focus as compared to narratives of Low Barrier personality types. All of the pictures were taken from the online photo management application http://www.flickr.com, and were publicised with 'no known restrictions on publication'.

2.4 Data

The assessment of inter-coder reliability and alternate-form reliability was based on 53 participants' open-ended responses in the Rorschach response task. This sub-sample was randomly selected from the full corpus (N = 526) based on Lacy and Riffle's (1996) suggestion that a sufficient subset for inter-coder reliability assessment should ideally not be less than 10 per cent of the full sample size.
The Rorschach responses in the sub-set (N = 53) had a total text length of 8,809 with a mean 166.21 of words per responses (SD = 106.41). The assessment of scoring consistency was based on the full data set (N = 526). The Rorschach responses (N = 526) had a total text length of 83,160 words with a mean of 158.10 words per response (SD = 96.43) and the picture response task had a text length of 277,997 words with a mean of 528.51 words per response (SD = 309.97). Narratives for everyday memories (N = 488) had a text length of 71,831 with a mean of 147.19 words per response (SD = 97.27) and narratives of dream memories (N = 450) had a text length of 62,005 with a mean of 137.79 words per response (SD = 125.16). Dream interpretations (N = 427) had a text length of 41,535 with a mean of 97.27 words per response (SD = 50.63).

All of the verbal responses were checked for correct spelling manually and spell-checked with the Microsoft Word Spelling and Grammar tool, through which typing errors (e.g. *batallion* for *battalion*) and incorrect first-letter capitalisations (e.g., *i* for *I*) were changed within the original texts. Due to the technical restrictions of the PROTAN content analysis software (Hogenraad, Daubies, Bestgen and Mahau 2003), brackets, hyphens and dashes were deleted from the corpus text. Apostrophes used in contractions (i.e. negations and personal pronouns with auxiliary verbs) were substituted with the original grammatical form, whereas apostrophes that marked a possessive case were deleted.

2.5 Content analysis

For the computerised content analysis, the BTD was applied to the texts using the PROTAN content analysis software program, which measures occurrences of category-based lexical content in texts (Hogenraad, Daubies, Bestgen and Mahau 2003). A lemmatisation process was then applied to reduce inflected words to their base forms. For example, *agrees*, *agreed*, *agreeing* were all reduced to *agree*. Subsequently, the lexical content of the segmented and reduced texts were matched against the predefined categories of the BTD.

The PROTAN computes two raw counts for the lexical occurrences. The density count shows how many distinct lexical items (i.e. types) match each dictionary category, whereas the frequency count represents how many lexical items in total (i.e. tokens) match the dictionary categories (Wilson 2008). For the purpose of this study, the frequency count measure was the most suitable for assessing inter-coder agreement, given that the frequency count represents an equivalent to the coders' manual frequency count for barrier and penetration imagery, which facilitates statistical comparisons. PROTAN also produces a density and frequency rate that takes segment length into account. Whereas the inter-rater coder reliability used the raw frequency counts for barrier, penetration and sum body boundary imagery, the alternate-form reliability and consistency of scoring of the BTD were assessed using a frequency rate that was calculated based on the following formula:

$$Frequency\ rate = \sqrt{\frac{frequency\ count}{no.\ of\ \ tokens\ in\ \ segment}}\, x1000$$

2.6 Statistical analysis of inter-coder reliability

Statistical calculations were performed with the statistical language and software of R (R Development Core Team 2011) using the kripp.alpha {irr} package (Garmer *et al.* 2012). Inter-coder reliability is assessed by calculating the agreement between the coders' annotations of the semantic items (Lombard *et al.* 2002). Although a variety of different coefficients have been suggested for assessing inter-coder agreement of nominal data (e.g. Percentage agreement, Cohen's kappa, Scott's pi, Spearman rho, Pearson r, etc.), there is not a single approach that represents the best statistical methodology, because every statistical procedure has strengths and weaknesses (Lombard *et al*. 2010). Krippendorffs's alpha (Krippendorff 2004) is the preferred method for measuring inter-coder agreement of linguistic data given that it is not based on nominal measures, i.e. ordinal, interval and ratio measures (Passonneau 2006). The linguistic

variables in this study were based on an ordinal measure. In particular, the alpha coefficient produces a more reliable agreement measure as compared to other coefficients. Hence, the coefficient generalises individual scores to reflect the reliability of the annotation procedure, which is independent of the individual scorers due to the exclusion of marginal disagreements from expected agreements. This procedure controls for differences in disagreement and expected agreement (Artstein and Poesio 2008: 17). The interpretation of Krippendorff's alpha assumes that correlation coefficients above $\alpha = .80$ are acceptable, whereas values below $\alpha = .80$ up to $\alpha = .67$ are difficult to interpret and may only allow researchers to make tentative conclusions (Fleiss 1981; Neundorf 2002; Krippendorf 2004). The alpha coefficient is calculated based on the following formula, in which D_O is the observed disagreement and D_E is the expected disagreement:

$$\alpha = 1 - \frac{D_O}{D_E}$$

This coefficient assumes two points of reference, which, in the absence of observed disagreement, becomes $D_O = O$ and $\alpha = 1$, thereby indicating perfect agreement. If the presence of observed agreement and disagreement is due to chance and expected disagreements are equal, then $D_O = D_E$ and $\alpha = O$, thereby indicating an absence of reliability (Krippendorff, 2004).

2.7 Inter-coder reliability procedure

Due to considerable variation in linguistic judgements across native English speakers but a lower frequency of this variation in educated native and non-native English speakers (Schmitt and Dunham 1999), it was deemed reasonable to invite university-related native, or near-native, English speakers to perform the manual coding, which should increase the general accuracy of the judgments regarding body boundary imagery. Two coders, one male native British English speaker and one male non-native British English speakers of near-native proficiency, were chosen, both of whom were undergraduate linguistics students.

The training process for the coders consisted of a briefing regarding the annotation task. Given that body boundary imagery represents a latent semantic variable that requires coders to use their subjective mental schemas, an initial pre-training session was conducted that involved a detailed, comprehensive explanation of the theoretical background of Fisher and Cleveland's (1956, 1958) body boundary concept and its lexical content classification scheme. Both coders were provided with a number of handouts outlining the theoretical basis

for and the coding scheme of body boundary imagery to familiarise themselves with the underlying theoretical and semantic contents of barrier and penetration imagery. A training session was scheduled for one week later, which involved an initial open discussion and clarification of the body boundary concept and its semantic classification. Coders were given some text samples to exercise the annotation of body boundary imagery. Once the coders felt familiar with the body boundary concept and coding scheme, a small sub-sample of the data was used to train the manual annotation of barrier and penetration imagery. The results were compared and discussed to assume an even 'calibration' between the coders, and any remaining questions and difficulties were clarified (Neuendorf 2002).

As proposed by Lombard and colleagues (2010), a separate study assessed both coders' annotation reliability using the manual annotation of barrier and penetration imagery with a small pilot sample (N = 10). This pilot sample was not included in the final study. The coders reviewed the barrier and penetration imagery independently without any help from the researcher. Coders annotated semantic units of the texts without being informed about the purpose and hypothesis of the study to reduce any possible confounding biases that could impact the validity of the results. Neuendorf (2002: 133) proposes that demand characteristics within the experimental situation (Orne 1962) (i.e. the tendency of research participants to produce responses that are assumed to be required by the researcher to confirm a particular hypothesis) might interfere with participants' freedom to produce responses that are independent of the researcher's influence. To create a new demand motivation that would counteract the tendency to comply with the demand characteristics of the experimental situation, the coders were told that they were not allowed to be informed about the experimental hypothesis of this study and that they should not try to determine the underlying theoretical construct of body boundary imagery within the narrower or wider framework of the research project (Rosenthal and Rosnow 1984). An initial pilot test of barrier and penetration imagery annotation indicated a high inter-coder agreement for barrier imagery (α = .95), penetration imagery (α = .93), and sum body boundary imagery (α = .95).

In particular, the coders were trained according to the semantic categories and lexical content of the BTD, as compared to Fisher and Cleveland's manual scoring scheme, such that coders were told to exclude polysemous words (e.g. *well*) and shelled sea animals (e.g. *Lobster Thermidor*). For the final coding, the coders were provided with a hardcopy of (N = 53) Rorschach responses from a sub-sample in order to independently and manually annotate the semantic items as barrier and penetration imagery. The researcher and the coders agreed that it

would take a two week period to complete the annotation task. Once the manually annotated texts were returned to the researcher, both coders were thanked for their participation and debriefed about the experimental purpose of this study. Subsequently, the researcher counted the manually annotated semantic items containing barrier and penetration imagery in both sub-samples, computed the sum frequency value for the barrier and penetration imagery scores, and computed the sum total of barrier boundary imagery scores.

2.8 Additional statistical analysis

Additional statistical calculations were performed using the statistical language and software from R (R Development Core Team 2011) and the R:commander {Rcmdr} package (Fox 2005). A Shapiro-Wilk test showed that the inter-coder sub-sample ($N = 53$), barrier and penetration imagery, $p < .01$, and sum body boundary imagery, $p < .05$, were not normally distributed in the Rorschach responses. In the complete data set ($N = 526$), barrier, penetration imagery and sum body boundary imagery were also not normally distributed in the experimental conditions (i.e. the Rorschach responses and picture response test responses, the narratives of everyday memories and dream memories, and dream interpretations; $p < .001$).

Thus, a non-parametric significance test appeared most suitable to assess the frequencies of barrier, penetration and sum body boundary imagery between the experimental conditions. A repeated measures Friedman test (Friedman 1937) was applied to the data with a post-hoc Wilcoxon signed rank test to compare the frequencies of barrier, penetration and sum body boundary imagery between the experimental conditions. A two-tailed non-parametric two-tailed Spearman's rank correlation coefficient (Spearman, 1904) was used to assess the alternate-form reliability and scoring consistency of barrier, penetration and sum body boundaries across the experimental conditions, as well as to provide an additional calculation of the inter-rater reliability assessment of body boundary imagery.

3 Results

3.1 Inter-coder reliability

The descriptive statistics for the manually coded barrier and penetration imagery are presented in Table 1. Although the sum body boundary imagery scoring did not differ substantially between coder 1 and coder 2, coder 1 showed a slightly higher coding of penetration imagery lexis and fewer barrier imagery lexis as compared to coder 2. A Krippendorff's alpha coefficient for ordinal data indi-

cated a sufficiently accurate inter-coder agreement of barrier imagery (α = .92), penetration imagery (α = .81), and sum body boundary imagery (α = .88). An additional series of Spearman's rank correlation coefficients also identified positive correlations between coder 1 and coder 2 for barrier imagery, = .94, $p < .001$, penetration imagery, $\rho = .84$, $p < .001$, and sum body boundary imagery, $\rho = .89$, $p < .001$. Thus, (H1) was maintained.

The acceptable alpha levels for barrier and penetration imagery indicate that both coders shared a good common-sense understanding of the BTD body boundary concept (i.e. barrier and penetration imagery), thereby demonstrating semantic validity of the semantic categories. The lack of a perfect agreement between coders might indicate that the semantic units in the semantic categories reflected a number of discrepancies in the overall application of the content analysis scheme. These discrepancies were related primarily to random annotation omissions of body boundary lexis, some degree of subjective interpretations of the body boundary concept, as well as the manual annotation of body boundary lexis that was not included in the BTD.

3.2 Alternate-form reliability

This part of the experiment assessed the alternate-form reliability of the BTD based on the assessment of whether manually scored frequencies for barrier and penetration imagery would be significantly correlated with computerised measures the same imagery, as measured in the Rorschach responses. The descriptive statistics showed that the means for barrier imagery was highest in coder 2, penetration imagery was highest in coder 1 as compared to coder 2 or the computerized scores, but for sum body boundary imagery was highest in the computerized scores than the manual scores (cf. Table 1).

Table 1: Descriptive statistics (mean, median, standard value and *interquartile* range) of coder 1 and 2, and computer-assisted coding of body boundary imagery

N = 53	***Variable***	***Mean***	***Median***	***SD***	***IQR***
Coder 1	***Barrier***	4.48	4.71	2.25	2.77
	Penetration	3.35	3.30	2.17	2.62
	Boundary sum	5.99	6.24	2.28	2.80
Coder 2	***Barrier***	4.83	4.94	1.99	2.58
	Penetration	2.90	3.13	1.74	1.67
	Boundary sum	5.93	5.80	1.87	2.13
BTD	***Barrier***	4.78	5.13	1.86	2.64
	Penetration	3.29	3.55	1.95	1.97
	Boundary sum	6.18	6.48	1.64	2.01

An analysis of the manually coded barrier and penetration imagery showed an overall moderately high positive correlation with the computerised frequency counts for the barrier, penetration, and sum body boundary imagery in the Rorschach responses (cf. Table 2). Overall, the moderately high effect size of the correlation coefficients between the manually and computerised coded lexis for the same response type suggest that there was acceptable inter-coder reliability. Both coders coded similarly in overall frequencies of barrier and penetration imagery, but the only moderately high correlation coefficients between manually coded lexis clearly indicated that coders differed in the annotation of individual lexical items. Consistent with (H2), the correlation coefficient effect size between manually and computerised coded lexis remained relatively moderately high when the manually coded variables were averaged (cf. Table 3).

Table 2: Spearman's rank correlation coefficients of coder 1 and 2, and computer-assisted coding of body boundary imagery

		Coder 1	*Coder 2*
Barrier	***Coder 1***	-	
	Coder 2	.889**	-
	BTD	.856**	.849**
Penetration	***Coder 1***	-	
	Coder 2	.860**	-
	BTD	.858**	.870**
Sum boundary	***Coder 1***	-	
	Coder 2	.800**	-
	BTD	.828**	.819**

Notes: * p < .05 level, ** p < .01 level

Table 3: Spearman's rank correlation coefficients of manual and computer-assisted coding of body boundary imagery

		Manual/BTD
Barrier	***1. Manually***	-
	2. BTD	.884**
Penetration	***1. Manually***	-
	2. BTD	.894**
Sum boundary	***1. Manually***	-
	2. BTD	.873**

Notes: * p < .05 level, ** p < .01 level

3.3 Consistency of scoring

The scoring consistency of the BTD was assessed by correlating barrier, penetration and sum body boundary imagery across the experimental conditions in the computer assisted scored of the full data set. The BTD would have high scoring consistency if a linguistic variable was significantly correlated with the same linguistic variable in any other experimental condition (e.g. Rorschach responses, picture response test, narratives of everyday and dream memories, and dream interpretations). The descriptive statistics show that the frequencies

of barrier imagery, penetration imagery, and sum body boundary were highest in the Rorschach responses and lowest in the dream interpretations (cf. Table 4).

Table 4: Descriptive statistics (mean, median, standard value and *interquartile* range) of computer-assisted coding of body boundary imagery in the experimental conditions

		Mean	*Median*	*SD*	*IQR*
Rorschach (BTD) (N = 526)	***Barrier***	4.83	4.86	1.92	2.16
	Penetration	2.75	3.04	2.06	4.13
	Sum boundary	5.93	5.92	1.91	2.21
Picture response test (BTD) (N = 526)	***Barrier***	4.22	4.22	1.15	1.45
	Penetration	1.88	2.03	1.13	1.26
	Sum boundary	4.76	4.77	1.11	1.44
Everyday memories (BTD) (N = 488)	***Barrier***	2.14	2.35	2.15	3.68
	Penetration	1.39	.00	1.85	2.80
	Sum boundary	3.03	3.24	2.31	4.71
Dream memories (BTD) (N = 450)	***Barrier***	3.22	3.63	2.42	4.94
	Penetration	1.43	.00	1.95	2.93
	Sum boundary	3.94	4.35	2.56	3.08
Dream interpretation (N = 427)	***Barrier***	1.66	.00	2.16	3.58
	Penetration	.72	.00	1.46	.00
	Sum boundary	2.17	2.31	2.32	4.15

A Friedman test indicated a significant difference in the frequency of barrier, penetration and sum body boundary imagery across the response types, $p < .001$. A post-hoc analysis with a pair-wise Wilcoxon signed-rank test tested for the significant difference between the medians of barrier, penetration and sum body boundary imagery across the experimental conditions (cf. Table 5).

Table 5: Wilcoxon signed-rank test results of body boundary imagery between experimental conditions

	Comparison	**Sig. level**
Barrier imagery	Rorschach > Picture response > Dreams > Everyday > Dream interpretation	**
Penetration imagery	Rorschach > Picture response > [Dreams = Everyday] > Dream interpretation	**
Sum boundary imagery	Rorschach > [Picture response = Dreams] > Everyday > Dream interpretation	**

A Spearman's rank correlation coefficient was applied to the data to assess the scoring consistency of the barrier, penetration and sum body boundary imagery across the experimental conditions (cf. Table 6). The results showed that barrier imagery in the Rorschach responses displayed a modest positive correlation with the picture response test, and barrier imagery also correlated positively between narratives of dream memories and dream interpretations. A positive correlation between Rorschach responses and the picture response test is also in accordance with other studies that identified correlations between Rorschach and TAT responses (e.g. Ackerman *et al*. 2001). Conversely, penetration imagery modestly correlated in the narratives for dream memories and dream interpretations only. Sum body boundary showed a modest positive correlation between the Rorschach responses and the picture response test, the picture response test and narratives of everyday memories, and sum body boundary also correlated between dream narratives and dream interpretations. The effect sizes in all correlations were low. The effect size, however, was higher in the positive correlation for barrier and sum body boundary imagery between dream memories and dream interpretations which might be related to the thematic similarity between both text types, for which most typically the dream interpretation would evaluate the recalled dream memory. Inconsistent with (H3), barrier, penetration and sum body boundary imagery reflect only a weak consistency of scoring across the experimental conditions.

Table 6: Spearman's rank correlation coefficients of computer-assisted coding body boundary imagery between experimental conditions

		1.	2.	3.	4.
Barrier	***1. Rorschach (BTD) (N = 526)***	-			
	2. Picture response test (BTD) (N = 526)	.135**	-		
	3. Everyday (BTD) (N = 450)	.074	.056	-	
	4. Dream (BTD) (N = 488)	.090	.003	-.032	-
	5. Dream interpretation (N = 427)	-.013	.064	-.062	.342**
Penetration	***1. Rorschach (BTD) (N = 526)***	-			
	2. Picture response test (BTD) (N = 526)	.012	-		
	3. Everyday (BTD) (N = 488)	.064	.079		
	4. Dream (BTD) (N = 450)	-.007	.008	.056	-
	5. Dream interpretation (N = 427)	.008	-.014	-.010	.320**
Sum	***1. Rorschach (BTD) (N = 526)***	-			
boundary	***2. Picture response test (BTD) (N = 526)***	.170**	-		
	3. Everyday (BTD) (N = 450)	.004	.101*	-	
	4. Everyday (BTD) (N = 488)	.073	-.055	-.031	-
	5. Dream interpretation (N = 427)	.015	.060	.018	.308**

Notes: * $p < .05$ level, ** $p < .01$ level

4 Discussion and conclusion

In summary, the results of this study demonstrated that the lexical content of the BTD reflects a reliable computer-assisted content analysis measure of body boundary imagery. The BTD yields quantitative data regarding barrier and penetration imagery frequencies that allows meaningful interpretations to be drawn. The first experiment indicated that the coders' judgments regarding classifying lexical content as barrier or penetration imagery showed an acceptable level of a shared common-sense understanding of the body boundary concept, and thus indicating also semantic validity. In addition, the results indicated acceptable alternate-form reliability in that the manually coded barrier and penetration imagery was highly correlated with the computer-assisted barrier and penetration scores. Conversely, the manual coding of independent coders revealed that the lexical content of the BTD could be improved by adding further semantic items, such as lexis relating to the clothing items.

The lack of correlations between barrier, penetration and sum body boundary imagery scores across the experimental conditions suggests a lack of scoring consistency at first glance. Thus, the concept of High and Low Barrier personality types as stable personality traits that are reflected through the consistent use of barrier imagery frequencies across all of the linguistic conditions appears to be challenged by the results of this study. The low scoring consistency of barrier and penetration imagery may be due to the relatively restricted lexical content of the body boundary categories, which are not always present in the content of a visual task interpretation (i.e. Rorschach response and picture response test) or in recalled autobiographical memory (i.e. narratives of everyday memories and dream memories). In fact, body boundary imagery represents only a small proportion of the overall words used in Rorschach responses (3.49%), in the picture response task (2.36%) in narratives of everyday memories (2.70%), in narratives of dream memories (4.54%), and in dream interpretations (l.03%). The restrictiveness of the body boundary lexical content was also evident in the narratives for dreams and in the dream interpretations. Although both text types are assumed to share at least some of the thematic of the recalled dream memory, the correlation coefficient effect size was only moderate, which provides some indication that lexical content might be context dependent (Schnurr *et al*. 1986).

Despite the statistically significant correlation coefficients between some of the experimental conditions, the small effect sizes indicate that the significant p-values might be related to the relatively large sample size. Thus, the small effect size shows a very low consistency of barrier, penetration, and body boundary imagery even between experimental conditions that were significantly different. In this sense, the small effect sizes identified in this study highlight the importance of effect size values as statistical measure to assess differences between the experimental and null hypotheses, rather than just reporting the obtained p-value (e.g. Michalczyk and Lewis 1980; Gigerenzer 2004). Low effect size represents an inherent and persistent statistical problem in content analysis research (Mergenthaler, personal communication), which may be related to the relatively short text samples in each of the experimental conditions. These short text samples may ultimately limit the probability of a body boundary lexis occurring when compared with other types of linguistic variables, whereas longer text samples would increase the probability of a more thematic diversity and vocabulary. Additionally, content words reflect only a small proportion of our usage-based vocabulary when compared to function words (such as pronouns, prepositions, articles, etc.) that provide an universal insight into quantitative views of social and psychological dimensions (see Argamon and Levitan 2005; Chang and Pennebaker 2007). Conversely, the low scoring consistency of body bound-

ary imagery might be also associated with variations in dedifferentiated cognition across the experimental conditions. Given that the Rorschach test requires high levels of free-associative thinking as compared to other experimental conditions, it might be that the frequencies of penetration are dependent on the level of regressive cognitive processes in the expected direction of primordial to conceptual thought functioning as put forward by Buck and Barden (1971); such a decrease has been however also identified in the frequencies of barrier imagery. Thus, the results showed that the Rorschach responses are relatively high in barrier and penetration imagery, the picture response task is relatively high in barrier and penetration imagery, everyday memories are relatively low in barrier and penetration imagery, dream memories are moderately high in barrier imagery but low in penetration imagery, and dream interpretations are low in barrier and penetration imagery. In this sense, the relationship between barrier and penetration imagery is not entirely transparent and it might perhaps be associated with the nature of the experimental conditions that can be differentiated between projective tests, and the recall and reflection of personal memories. The results of this study might then confirm that barrier and penetration imagery reflect related personality dimensions as compared to opposite ends of a polar personality model (Fisher and Cleveland 1956, 1958). The results of this study provide also some support to Wilson's (2009) assumption that penetration imagery would be related to context dependent regressive cognitive functioning, whereas an increase of barrier imagery might represent a compensatory function of an enduring uncertain body boundary awareness associated with low barrier personality as "they serve, in a real sense, as barriers which differentiate the self from the other" (2009: 13).

An interesting methodological detail in this study was the use of online administration of the Rorschach inkblot test. Body boundary imagery in verbal Rorschach responses have been typically only used to the assessment of High and Low Barrier personality. Conversely, the primary purpose of this study was to reach a wide population to participate in the survey and thus to obtain a large sample size of various experimental conditions as a means to assess the reliability of the lexical categories of the BTD. Based on this premise, the use of an online-based Rorschach test might not represent a methodological issue taking into consideration that the use of online-based psychological assessment has been also associated with some advantages over face-to-face personality testing situations, such as increased disclosure (Buchanan 2002).

Overall, the results of this study were satisfactory in that they provided acceptable levels of inter-coder reliability and alternate-form reliability. The low consistency of scoring might indicate that body boundary awareness might not

necessarily represent a stable personality trait as put forward by Fisher and Cleveland (1956, 1958), but instead, it might be dependent on the level of cognitive dedifferentiation. Thus, future research should investigate further the relationship between body boundary imagery and level of dedifferentiated cognition.

Acknowledgments: I would like to express my gratitude to Bart Modderkolk and David de Winne for their research assistance, and also to my supervisor Dr. Andrew Wilson.

Notes

1. Wilson (2006) excluded the lexical items *boot(s)*, *Wellington(s)*, *welly/wellies*, and *mud* to control for increased lexical focus on boots in the rubber boot fetish narratives. In fact, the first version of Fisher and Cleveland's body boundary scoring system (1956) contained *clothing items with unusual covering and decorative function*, and only *buildings with unusual structures*, whereas the second edition (1958) included all types of *clothing items*, *vehicles*, and *buildings*.
2. The abbreviation M stands for *mean* which indicates the statistical average value, and SD stands for *standard deviation* which indicates the statistical variability of the average value.

References

Ackerman, Steven J., Mark J. Hilsenroth, Amanda J. Clemence, Robin Weatherill and I. Christopher Fowler. 2001. Convergent validity of Rorschach and TAT scales of object relations. *Journal of Personality Assessment* 77: 295–306.

Anzieu, Didier. 1989. *The skin ego*. New Haven: Yale University Press.

Argamon, Shlomo and Shlomo Levitan. 2005. Measuring the usefulness of function words for authorship attribution. *Proceedings of the 2005 ACH/ALLC conference*, June 2005. Victoria, Canada.

Artstein, Ron and Massimo Poesio. 2007. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34: 555–596.

Bick, Esther. 1968. Experience of the skin in early object relations. *International Journal of Psycho-Analysis* 49: 484–486.

Buchanan, Tom. 2002. One assessment: Desirable or dangerous? *Professional Psychology: Research and Practice* 33: 148–154.

Buck, Lucien A. and Michael Barden. 1971. Body image scores and varieties of consciousness. *Journal of Personality Assessment* 35: 309–314.

Burris, Christopher T. and John K. Rempel. 2004. 'It's the end of the world as we know it': Threat and the spatial-symbolic self. *Journal of Personality and Social Psychology* 86: 19–42.

Cariola, Laura A. 2012. A case study of primary process language and body boundary imagery in discourses of religious-mystical and psychotic altered states of consciousness. *Empirical Text and Cultural Research – ETC* 5: 36–61.

Chang, Cindy and James W. Pennebaker. 2007. The psychology of function words. In K. Fiedler (ed.). *Social communicatio*n, 343–359. New York: Psychology Press.

Fisher, Seymour and Sidney E. Cleveland. 1956. Body-image boundaries and style of life. *Journal of Abnormal and Social Psychology* 52: 373–379.

Fisher, Seymour and Sidney E. Cleveland. 1958. *Body image and personality*. New York: Dover Publications.

Fisher, Seymour. 1986. *Development and structure of the body image*. Hillsdale: Lawrence Erlbaum.

Fleiss, Joseph L. 1981. *Statistical methods for rates and proportions*. New York: John Wiley.

Fox, John. 2005. The R commander: A basic statistics graphical user interface to R. *Journal of Statistical Software* 14: 1–42.

Friedman, Milton. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32: 675–701.

Garmer, Matthias, Jim Lemon, Ian Fellows and Suspendra Singh. 2012. *Various coefficients of interrater reliability and agreement*. Available at http://www.cran.r-project.org/web/packages/irr/irr.pdf. Last accessed on 19 January, 2014.

Gigerenzer, Gerd. 2004. Mindless statistics. *The Journal of Socio-Economics* 33: 587–606.

Guimon, Jose. 1997. Corporality and psychoses. In J. Guimon (ed.). *The body in psychotherapy*, 63–72. Basel: Karger.

Hogenraad, Robert, Claude Daubies, Yves Bestgen and Pierre Mahau. 2003. *Une théorie et une méthode générale d'analyse textuelle assistée par ordinateur: Le systéme PROTAN (PROTocol ANalyzer)*. 32-bits version of November 10, 2003 by Pierre Mahau. Louvain-la-Neuve: Psychology Department, Catholic University of Louvain.

Jackson, Sherri. 2011. *Research methods and statistics: A critical thinking approach*. Belmont: Wadsworth.

Kolbe, Richard H. and Melissa S. Burnett. 1991. Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research* 18: 243–250.

Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology.* Thousand Oaks: Sage. (First published 1980).

Lacy, Stephen and Daniel Riffle. 1996. Sampling error and selecting intercoder reliability samples for nominal content categories: Sins of omission and commission in mass communication quantitative research. *Journalism & Mass Communication Quarterly* 73: 969–973.

Lombard, Matthew, Jennifer Snyder-Duch and Cheryl Campanella Bracken. 2002. Content analysis in mass communication: Assessment and reporting of inter-coder reliability. *Human Communication Research* 28: 587–604.

Lombard, Matthew, Jennifer Snyder-Duch and Cheryl Bracken. 2010. *Practical resources for assessing and reporting intercoder reliability in content analysis research projects.* Available at http://matthewlombard.com/reliability/. Last accessed on 29 January, 2014.

Michalczyk, Alan E. and Lloyd A. Lewis. 1980. Significance alone is not enough. *Journal of Medical Education* 55: 835–838.

Morgan, Christina D. and Henry A. Murray. 1935. A method of investigating fantasies: The Thematic Apperception Test. *Archives of Neurology and Psychiatry* 34: 289–306.

Neuendorf, Kimberly A. 2002. *The content analysis guidebook*. London: Sage Publications.

Neundorf, Kimberly A. and Paul D. Skalsi. 2010. Extending the utility of content analysis via the scientific method. Manuscript in support of presentation to the Social Science Computing Workshop, University of Hawaii, Honolulu. Available at http://www.manoa.hawaii.edu/ccpv/workshops/KimberlyNeuendorf.pdf.

Ogden, Thomas H. 1989. *The primitive edge of experience*. Northvale: Jason Aronson.

O'Neill, Regina M. 2005. Body image, body boundary and the Barrier and Penetration Rorschach scoring system. In R. F. Bornstein and J. M. Masling (eds.). *Scoring the Rorschach: Seven validated systems*, 159–189. London: Lawrence Erlbaum.

Orne, Martin T. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist* 17: 776–783.

Passonneau, Rebecca. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Genoa, Italy.

Potter, W. James and Deborah Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research* 27: 258–284.

R Development Core Team. 2011. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at http://www.R-project.org. Last accessed on 19 January, 2014.

Rorschach, Hermann 1921. *Psychodiagnostik*. Leipzig: Ernst Bircher Verlag.

Rosenthal, Robert and Ralph L. Rosnow. 1984. *Essentials of behavioural research: Methods and data analysis*. London: McGraw-Hill.

Rourke, Liam, Terry Anderson, D. R. Garrison and Walter Archer. 2000. Methodology issues in the content analysis of computer conference transcripts. *International Journal of Artificial Intelligence in Education* 11: 8–22.

Saraceni, Carlo, Giuseppe Ruggeri and D. Filocamo. 1980. Studio sperimentale con il test di Rorschach sulle modificazioni dell'immagine corporea in ipnosi. *Archivio di Psicologia, Neurologia e Psichiatria* 41: 50–64.

Schmeidler, Gertrude and Lawrence LeShan. 1970. An aspect of body image related to ESP scores. *Journal of the American Society for Psychical Research* 64: 211–218.

Schmitt, Norbert and Bruce Dunham. 1999. Exploring native and non-native intuitions of Word frequency. *Second Language Research* 15: 389–411.

Schnurr, Paula P., Stanley D. Rosenberg, Thomas E. Oxman and Gary J. Tucker. 1986. A methodological note on content analysis: Estimates of reliability. *Journal of Personality Assessment* 50: 601–609.

Singletary, Michael W. 1994. *Mass communication research: Contemporary methods and applications*. New York: Longman.

Spearman, Charles. 1904. The proof and measurement of association between two things. *American Journal of Psychology* 15: 72–101.

Tinsley, Howard. E. A. and David J. Weiss. 1975. Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology* 22: 358–376.

Tustin, Frances. 1981. *Autistic states in children*. London: Routledge.

Weber, Robert P. 1990. *Basic content analysis*. London: Sage.

West, Alan N. 1991. Primary process content in the King James Bible: The five stages of Christian mysticism. *Computers and the Humanities* 25: 227–238.

Wilson, Andrew. 2006. The development and application of a content analysis dictionary for body boundary research. *Literary and Linguistic Computing* 21: 105–110.

Wilson, Andrew. 2008. *Psychosomatic cycles and the liturgical year: A case study and framework for research*. Gottingen: Cuvillier Verlag.

Wilson, Andrew. 2009. Barrier and penetration imagery in altered states of consciousness discourse: Replicating the five-stage model of Christian mysticism in the Bible. In W. Oleksy and P. Stalmaszczyk (eds.). *Cognitive approaches to language and linguistic data: Studies in honor of Barbara Lewandowska-Tomaszczyk* (Polish Studies in English Language and Literature 27), 357–372. Frankfurt am Main: Peter Lang.

Wilson, Andrew. 2011. The regressive imagery dictionary: A test of its concurrent validity in English, German, Latin and Portuguese. *Literary and Linguistic Computing* 26: 125–135.

Appendices

Appendix 1

Semantic categories and examples of barrier and penetration imagery in the BTD (Wilson 2006), including all clothing items, vehicles and buildings

Barrier imagery	**Examples of semantic items**
Clothing items	*Dress, robe, costume*
Animal with distinctive or unusual skins, including shelled creatures	*Alligator, badger, peacock, snails, shrimp*
Enclosed openings in the earth	*Valley, ravine, canal*
Unusual animal containers	*Bloated, kangaroo, pregnant*
Overhanging or protective surfaces	*Umbrella, dome, shield*
Armoured objects or objects dependent on their own walls	*Armour, battleship, ship*
Things being covered, surrounded or concealed	*Covered, hidden, behind*
Buildings	*Bungalow, cathedral, tower (except building that relate to social institutions, e.g. church, hospital, school.*
Enclosed vehicles	*Car, ship, truck*
Things with unusual container like shapes or properties	*Bagpipes, chair, throne*
Unique structures	*Tent, fort, hut*
Miscellaneous barrier words	*Basket, bubble, cage*
Penetration imagery	
Reference to the mouth being opened or used for intake or expulsion	*Eating, tongue, yawning*
Reference to evading, or bypassing or penetrating through the exterior of an object	*Autopsy, fluoroscope, x-ray,*
References to the body wall being broken, fractured, injured and damaged, including degeneration of surfaces	*Bleeding, stabbed, wounded, withered*
Openings in the earth that have no set boundaries	*Abyss, fountain, geyser*
All openings	*Anus, doorway, entrance*
Things which are insubstantial and without palpable boundaries	*Ghost, mud, shadow*
Transparency	*Crystal, see-through, transparent*
Miscellaneous penetration words	*Broken, frayed, hole*

Appendix 2

Figure 1: Picture 4 of picture response test http://www.flickr.com/photos/powerhouse_museum/3640355880/

Figure 2: Picture 4 of picture response test http://www.flickr.com/photos/osucommons/5139906857/

Figure 3: Picture 4 of picture response test http://www.flickr.com/photos/statelibrary-ofnsw/3294694544/

Figure 4: Picture 4 of picture response test http://www.flickr.com/photos/statelibraryqueensland/4292454948/