The dimensionality of inference making: Are local and global inferences distinguishable?

Language and Reading Consortium

and

Marloes M. L. Muijselaar

Lancaster University, UK

University of Amsterdam, Netherlands

Accepted for publication in Scientific Studies of Reading, 13 August 2017

Author Note:

This paper was prepared by a Task Force of the Language and Reading Research Consortium (LARRC) consisting of Kate Cain (Convener) and Marloes M. L. Muijselaar. LARRC project sites and investigators are as follows:

Ohio State University (Columbus, OH): Laura M. Justice (Site PI), Richard Lomax, Ann O'Connell, Jill Pentimonti[1], Stephen A. Petrill[2], Shayne B. Piasta

Arizona State University (Tempe, AZ): Shelley Gray (Site PI), Maria Adelaida Restrepo.

Lancaster University (Lancaster, UK): Kate Cain (Site PI).

University of Kansas (Lawrence, KS): Hugh Catts[3] (Site PI), Mindy Bridges, Diane Nielsen.

University of Nebraska-Lincoln (Lincoln, NE): Tiffany Hogan (Site PI), Jim Bovaird, J. Ron Nelson.[4]

MGH Institute of Health Professions (Boston, MA): Tiffany Hogan (Site PI)

1. Jill Pentimonti is now at American Institutes for Research (AIR).

2. Stephen A. Petrill was a LARRC co-investigator from 2010-2013.

3. Hugh Catts is now at Florida State University.

4. J. Ron Nelson was a LARRC co-investigator from 2010-2012.

This work was supported by grant # R305F100002 of the Institute of Education Sciences' Reading for Understanding Initiative. We are deeply grateful to the numerous staff, research associates, school administrators, teachers, children, and families who participated. Key personnel at study sites include: Crystle Alonzo, Lisa Baldwin-Skinner, Lauren Barnes, Garey Berry, Beau Bevens, Jennifer Bostic, Shara Brinkley, Janet Capps, Tracy Centanni, Beth Chandler, Lori Chleborad, Emmanuel Cortez, Willa Cree, Dawn Davis, Kelsey Dickerhoof, Jaclyn Dynia, Michel Eltschinger, Kelly Farquharson, Tamarine Foreman, Yvonne Fraser, Abraham Aldaco Gastelum, Rashaun Geter, Sara Gilliam, Alexandria Hamilton, Cindy Honnens, Miki Herman, Hui Jiang, Elaine Joy, Jaime Kubik, Trudy Kuo, Gustavo Lujan, Chi Luu, Junko Maekawa, Carol Mesa, Denise Meyer, Maria Moratto, Kimberly Murphy, Marcie Mutters, Amy Pratt, Trevor Rey, Lizeth Sanchez-Verduzco, Amber Sherman, Shannon Tierney, Stephanie Williams, and Gloria Yeomans-Maldonado.

The views presented in this work do not represent those of the federal government, nor do they endorse any products or findings presented herein. Correspondence concerning this work should be sent to Kate Cain, k.cain@lancaster.ac.uk, Department of Psychology, Lancaster University, LA1 4YF, UK.

Abstract

We investigated the dimensionality of inference making in samples of 4- to 9-year-olds (*N*s = 416 - 783) to determine if local and global coherence inferences could be distinguished. Additionally, we examined the validity of our experimenter-developed inference measure by comparing with three additional measures of listening comprehension. Multi-trait, multi-method modeling determined that the best fitting model included both text and inference factors, but the factor loadings of these final models showed that local and global inference factors could not be measured reliably. The Inference Task as a whole was reliable, and also showed good validity at all grade levels.

Keywords: inference making, validity, dimensionality, listening comprehension

The dimensionality of inference making: Are local and global coherence inferences distinguishable?

To comprehend written or spoken text, a coherent mental representation of the text's meaning is constructed (Kintsch & van Dijk, 1978). Since most texts do not explicitly provide all of the essential information to establish coherence, inference making is necessary to establish both local and global coherence (Graesser, Singer, & Trabasso, 1994). Inferences draw on information in the text, as well as information outside of the text, such as vocabulary and background knowledge. Local coherence inferences are necessary in order to integrate information from adjacent pieces of text, whereas global coherence inferences are used to fill in details not explicitly stated that are needed to construct a globally coherent representation of text meaning, for example inferences about themes, morals, and settings (Currie & Cain, 2015; Cain & Oakhill, 1999; 2014; Freed & Cain, 2017; Long & Chong, 2001). Inference making in general is critical to successful reading and listening comprehension in children both concurrently and longitudinally, over and above cognitive factors such as general ability and memory (Oakhill & Cain, 2012; Cain, Oakhill, & Bryant, 2004; Florit, Roch, & Levorato, 2014; Kim, 2016).

Local and global coherence inferences both support the construction of a model of a text's meaning, but serve different functions, as noted above; a text can be locally coherent without being globally coherent (Graesser et al., 1994). Thus, a valid research question is to ask whether or not these two aspects of inference making are distinguishable. There is some empirical evidence to suggest this might be the case. First, for poor comprehenders, global coherence inference making is more greatly impaired than local coherence inference making when compared to peers, for both children (Cain, 1999) and adults (Long & Chong, 2001). Second, for both reading (Cain & Oakhill, 2014) and listening (Currie & Cain, 2015) comprehension, vocabulary is a stronger predictor of global than local coherence inference making. In addition, working memory is more strongly related to global than local coherence inference making for both written (Chrysochoou, Bablekou, & Tsigilis, 2011) and spoken (Currie & Cain, 2015) texts. These findings reflect the strong relation between listening and reading comprehension, once the contribution of word reading to the latter has been taken into account (Language and Reading Research Consortium, 2015). In sum, previous research indicates that local and global coherence inferences, which serve different functions in a text, show different strengths of relations with language and cognitive variables.

As described above, inferences that support these two aspects of coherence (local and global) have been distinguished theoretically (Graesser et al., 1994; Long & Chong, 2001) and also empirically in terms of their predictors for both written and spoken texts (Cain & Oakhill, 2014; Chrysochoou et al., 2011; Currie & Cain, 2015). However, research to date has not examined if local and global coherence inferences can be *statistically* distinguished. In other words, the dimensionality of inference making has not been investigated. In contrast, the existence of specific language subskills of reading comprehension, in addition to word reading ability, was examined in some of the very early studies in the field of reading research using confirmatory and exploratory factor analyses (e.g., Davis, 1944; Spearritt, 1972; Thorndike, 1973). These studies largely support the viewpoint that reading comprehension is a one-dimensional construct and that specific reading comprehension skills could not be measured reliably.

More recently, two studies have investigated the dimensionality of reading comprehension using more sophisticated data analytic approaches (Basaraba, Yovanoff, Alonzo, & Tindal, 2013; Muijselaar et al., 2017). These studies report contradictory findings. The analysis of Basaraba et al. (2013) suggests that reading comprehension is a multi-dimensional construct; specifically, this showed that literal, inferential, and evaluative questions could be distinguished with a bifactor analysis on an item pool with 20 questions originating from one text. In contrast, Muijselaar et al. (2017), using multi-trait, multi-method modeling (MTMM) on an item pool of 77 questions originating from different texts, found that specific reading comprehension text types and question types could not be measured reliably and thus may not be separable constructs. Note that neither study distinguished different types of inference.

The Current Study

In sum, to our knowledge, previous research has not examined whether items that are specifically constructed to measure the ability to generate inferences that serve two different coherence functions (establishing local and global coherence) can be distinguished. In this study, we examined the structure of inference making using a bespoke measure of inference making informed by previous research on children's inference making (e.g., Cain & Oakhill, 1999; 2014; Currie & Cain, 2015; Freed & Cain, 2017). Our study started with children in preschool and, thus, assessed inference making from spoken texts. There were two aims. The first was to determine if local and global inferences can be distinguished and measured reliably. The second was to assess the validity of this measure in relation to standardized assessments of general listening comprehension; as noted, previous work has established the importance of inference making to reading comprehension, but its relation to listening comprehension in young children, including pre-readers, has not been determined to date.

In our study, we wanted to use a more sophisticated data analytic approach than the confirmatory and exploratory factor analyses used in the very early studies on the dimensionality of reading comprehension (e.g., Davis, 1944; Spearritt, 1972; Thorndike, 1973), namely a hierarchical data analytic approach that has often been used to test the dimensionality of intelligence (e.g., Carroll, 2003; Gustafsson, 1984; 2002; Undheim & Gustafsson, 1987). In these studies, a second-order factor model was used (Gustafsson 1984; 2002) in which the second-order factor (i.e., general intelligence) represented the relations between the first-order factors (i.e. verbal and nonverbal), and the first-order factors represented the relations between the corresponding subtests (see for an example Figure 1). A disadvantage of second-order factor models is that the variance explained by the second-order factor and the first-order factors cannot be distinguished. A bifactor model is a specific type of second-order factor models (e.g., Chen, West, & Sousa, 2006; Gustafsson, & Åberg-Bengtsson, 2010; Schmid & Leiman, 1957; see Figure 1 for an example), in which the general factor represents the variance that all subtests have in common and the specific factors depict the variance that is explained by groups of subtests, given the variance explained by the general factor. In such a bifactor model, the variance explained by general and specific factors can be interpreted separately. Specifically, in item bifactor models, each item is represented by a general factor, and one of the specific factors (Cai, Yang, Hansen, 2011; Gibbons et al., 2007; Gibbons & Hedeker, 1992; Undheim & Gustaffson, 1987).

In the current study, we used a complex MTMM model on the item level in which such a bifactor model is incorporated (Eid et al., 2008; Maul, 2013; see Figure 2, Model 3), because we also wanted to take into account the fact that our items were nested within texts. In this model, a trait represents a construct that is measured with different tests, and a method factor refers to the variance that these different tests have in common because they use the same method of measurement (Little, 2013). In the MTMM model in the current study, the texts in which the questions were nested were the method factors, and the local and global inferences were represented by the specific trait factors. The relations between all items are represented by the general trait, which was the general inference making factor.

Method

Participants

Children were participants in a larger longitudinal study on listening and reading comprehension conducted by the Language and Reading Research Consortium as part of the

IES funded Reading for Understanding programme, in which children from pre-kindergarten (P), kindergarten (K), grade 1 (G1), grade 2 (G2), and grade 3 (G3) were followed in consecutive years with the final testing point for each being grade 3. Therefore, children who were in pre-kindergarten in Year 1 of the study were followed for 5 years, whereas those in grade 3 at the start of the study were tested just once. We combined our datasets by grade, such that children in pre-kindergarten at the start of the study contributed data to the analysis of each grade level, and those in grade 3 contributed data only to the grade 3 analysis. For children who repeated a grade, we only used the data from the first year. For our analysis, we had the following numbers of participants by grade: 416 pre-kindergartners (241 boys, *M* = 5 years and 1 month, *SD* = 4.33 months), 520 kindergartners (289 boys, *M* = 6 years and 1 month, *SD* = 3.93 months), 620 first graders (324 boys, *M* = 7 years and 1 month, *SD* = 4.10 months), 724 second graders (380 boys, *M* = 8 years and 1 month, *SD* = 4.19 months), and 783 third graders (400 boys, *M* = 9 years and 1 month, *SD* = 4.10 months). The geographic locations of our sample resulted in a sample that was less racially and ethnically diverse than the general U.S. population. Our sample did reflect a wide range of income levels (12.8% < 30K, 25.3% 31-60K, 61.9% > 60K), with 14.6% on Free/Reduced Lunch. For further information about the study and participants, see Language and Reading Research Consortium (LARRC), Farquharson, and Murphy (2016).

Materials

For the present analysis, we present data from an experimenter designed measure of inference making, which was presented aurally, and three additional measures of general listening comprehension.

Inference making task. This comprised two stories at each grade level, each one followed by eight questions to assess the ability to generate local and global coherence inferences (four questions each for local and global coherence inferences per text). The stories and questions were based on the work of Cain and Oakhill (1999) and Oakhill and Cain (2014). Other work using these texts demonstrates strong inter-rater discrimination between the two question types (Currie & Cain, 2015). The second story at each grade level was repeated at the subsequent grade, such that there was one unique story at each grade. A full set of stories together with examples of acceptable responses is presented in Appendix A. Questions were scored as either correct, partially correct, or incorrect (0 – 2 points) using a rubric; awarding credit for partially correct responses is shown to be sensitive to detecting inference making ability in beginner readers (Paris & Paris, 2005; Silva & Cain, 2012). The average score on the local and global questions, and the average score on all questions were used in the analyses (range of 0 – 2 points). The reliability of the test at each grade level was acceptable (Cronbach's alphas of .78, .64, .71, .74, and .69 for P, K, and grades 1 to 3 respectively). The reliabilities for the composite scores of the local and global inferences calculated across stories were lower ranging from .44 to .69.

Listening comprehension. Three measures of discourse-level listening comprehension were administered. The Listening Comprehension Measure (LCM) was adapted from the Qualitative Reading Inventory – Fifth Edition (QRI-5; Leslie & Caldwell, 2011). Children were required to listen to narrative and expository paragraphs read aloud and to respond to several inferential and non-inferential questions after each one. Different passages were administered at each grade: there were 3 stories for P, K, and G1, and 4 stories for G2 and G3. The number of questions after each story differed: Children were asked between 14 and 32 questions in total. Cronbach's alphas for our sample were adequate to good across grades (.78, .79, .65, .75, and .83, from P to grade 3, respectively). A modified version of the CELF 4 Subtest Understanding Spoken Paragraphs (USP; Semel, Wiig, & Secord, 2003) was administered to evaluate children's ability to understand oral narratives. The modification was to present two of the three passages for kindergarten through to grade

3, to reduce the administration time. The CELF does not have materials for pre-kindergarten, so we constructed one new story and questions for this age group, which was administered along with one of the original kindergarten stories. There were five open-ended questions after each text. Cronbach's alphas were low to acceptable (-.08 to .71 for the different grades). Despite this low internal consistency, we retained the CELF in our analyses for the following reasons. First, other work has shown that a latent construct of listening comprehension including this modified CELF as an indicator has good construct reliability, calculated using Hancock's H (Language and Reading Research Consortium, 2015). The inclusion of the modified CELF subtest in this latent construct enables integration of findings across studies using the same latent construct. The Test of Narrative Language – Receptive (TNL; Gillam & Pearson, 2004), children listened to three passages that were read aloud and responded to between 9 and 11 open-ended questions after each one. The same passages were used at each grade. Internal consistency was moderate to good for each grade: .87, .87, .69, .73, and .53, for increasing grade). For the LCM, USP, and TNL, the total raw scores were utilized in the analyses.

Procedure

Children were tested over the course of multiple sessions within a 5- or 6-month time frame (January to May/June) each year. The assessments were administered in blocks, with breaks in between individual assessments and also between blocks where these were administered on the same day. Each block lasted no longer than 60 minutes. The measures were administered in a quiet room within the child's school, local university site, community center or home by trained research staff.

Analysis plan

The dimensionality of the Inference Task was investigated with several confirmatory factor models. As a first step, a one-factor model was created in which all items loaded on a general inference making factor (see Figure 2, Model 1). In the second step, a bifactor model was estimated in which all items loaded on a general inference making factor, and in addition, on the text to which they belonged (see Figure 2, Model 2). In the third step, a MTMM model was estimated (see Figure 2, Model 3), in which each item loaded on a local or global inference factor (based on the coding used in the development of the measure), in addition to the loadings on the general factor and one of the text factors. The latent factors in all models were specified to be uncorrelated. Thus, in the MTMM model, each item is described by its relation with the general inference making factor, with one of text factors, and one of the inference factors (local or global inference).

The analyses were carried out with Mplus Version 7.11 (Muthén & Muthén, 2012). Since the items were dichotomous, WLSMV (robust weighted least squares estimation) was used to obtain parameter estimates and therefore, theta-parameterization was used. The fit of the models was evaluated with inspection of three indices: the chi-square goodness-of-fit test-statistic, the RMSEA, and the CFI (Kline, 2011). A nonsignificant chi-square indicated good overall model fit, whereas a significant chi-square showed poor fit. However, in combination with relatively large sample sizes, the chi-square statistic often is significant. Therefore, the ratio χ^2/df was also used to evaluate model fit. A χ^2/df ratio < 2 confirmed a good fit (Maruyama, 1998). A model with an RMSEA below .05 has a good approximate fit, an RMSEA between .05 and .08 was taken as satisfactory approximate fit, and values above .10 indicated poor approximate model fit (Browne & Cudeck, 1993). A model with a CFI larger than .95 had a good incremental fit to the data, and a CFI larger than .90 was taken as acceptable (Hu & Bentler, 1999). Differences between nested models were tested with the corrected chi-square difference test (with Satorra-Bentler correction; DIFFTEST command in Mplus) (Kline, 2011), since the difference between two nested models with WLSMV as estimator does not have a chi-square distribution. In addition, the factor loadings were inspected and the variance explained and the reliability scores were calculated to evaluate the local fit of the final models. To calculate the variance explained by the latent factors of the final model, the following formula was used: $R^2 = (\Sigma\lambda_i) / 16$. In this formula, $\Sigma\lambda_i$ represents the sum of the standardized factor loadings of all items on a specific latent factor, and 16 is the total number of items. The reliability of the latent factors was computed with: $\rho_c = (\Sigma\lambda_i)^2 / ((\Sigma\lambda_i)^2 + \Sigma\theta_{\varepsilon i})$ (Brown, 1989). In this formula, $\theta_{\varepsilon i}$ represents the standardized residual variance of an item, which is calculated with the following formula: $\theta_{\varepsilon i} = 1 - \lambda_i^2$.

As an additional test of reliability, we exploited the fact that our samples were re-tested each year (up until grade 3). We computed Pearson correlations between the total local and the total global coherence inference scores obtained in consecutive years. In addition, we computed Pearson correlations between the total raw score of the Inference Task between all consecutive years.

Next, we investigated the validity of the Inference Task. First, the correlations of the inference making test with the total scores of the different listening comprehension measures were examined. In addition, we examined the correlation between the inference making test and a factor score of the listening comprehension measures to get a more reliable estimate of the true correlation of the Inference Task with the listening comprehension construct.

Results

Data Screening and Descriptive Statistics

Data were checked for outliers and missing values. Outliers (data points: $-3 < z < 3$) were removed from the datasets. Participants with missing data for all inference making questions were removed from the dataset. As a result, data from 386 pre-kindergartners, 491 kindergartners, 577 first graders, 678 second graders, and 690 third graders, were included in the analyses. From the composite scores of these final datasets, 10%, 4%, 4%, 3%, and 3% of the data was missing for P, K, and Grades 1 to 3, respectively. Less than 5% of missing data is not a problem in further analyses (Tabachnick & Fidell, 2013), and the data in P were missing completely at random (Little's MCAR test $p > .05$). Descriptive statistics for the measures for listening comprehension are displayed in Table 1. Correlations among those measures are presented in Table 2.

The Dimensionality of Inference Making

Several confirmatory factor models were estimated to investigate the dimensionality of inference making in each grade. For pre-kindergarten, a one-factor model had a significant chi-square value, which indicates poor overall model fit, but a chi-square/*df* ratio < 2 that indicated good model fit. This one-factor model had a good approximate fit to the data and an acceptable incremental fit (see Table 3). In a second step, two latent factors for the two texts were added. The second model gave estimation problems, since the loading of one item on a specific latent text factor was highly negative. This problem was solved by fixing this factor loading at 0. The bifactor model had a better fit than the one-factor model (see Table 3, P, model 1 vs 2): it had a good overall model fit based on the chi-square/*df* ratio, and a good approximate and incremental model fit. In a third step, two latent factors for the two aspects of coherence inference making were added. This MTMM model had a better fit to the data than the bifactor model (see Table 3, P, model 2 vs 3): the overall model fit, the approximate fit, and the incremental model fit were good. This MTMM model was therefore chosen as the final model (see Figure 2, Model 3 for an illustration of the final model).

The modeling process for each successive grade showed a very similar pattern. For children in kindergarten, the one-factor model had a good overall fit to the data, a good approximate fit, and an acceptable incremental fit. This model was improved by adding the latent text factors (see Table 3, K, model 1 vs 2) and this bifactor model was further improved by adding the latent inference factors (see Table 3, K, model 2 vs 3). For grade 1, the one-factor model had a poor overall and incremental fit, but a good approximate fit to the data and for grade 2 the one-factor model had a poor overall model fit, a good approximate fit, and an acceptable incremental fit. For both grades, the bifactor model had a better fit than the one-factor model (see Table 3, grades 1 and 2, model 1 vs 2) and the best fitting model included additionally the two latent inference factors (see Table 3, grades 1 and 2, model 2 vs 3). The final MTMM model for kindergarten, grade 1 and grade 2 each had a good overall, approximate and incremental model fit.

Lastly, the one-factor model for grade 3 gave some estimation problems, because the factor loading of one specific item on the general factor was very high. This problem was solved by fixing the factor loading of this item on the general factor at .9, and fixing the corresponding residual variance at .19. This model had a poor overall model fit, a satisfactory approximate fit, and an acceptable incremental fit. The bifactor model fitted the data better than the one-factor model (see Table 3, grade 3, model 1 vs 2), and the MTMM model was again taken as the final model since this model had the best fit to the data (see Table 3, grade 3, model 2 vs 3) and had good overall, approximate and incremental fit.

Interpretation of the Factor Loadings and Reliability Scores of the Final Models

For all grades, the MTMM model was chosen as the final model. From these final models, the median, minimum, and maximum factor loadings for each latent factor are reported in Table 4. The factor loadings of the items on the latent text and inference factors were very small, or sometimes even negative. For example, for the prekindergarten sample, one question in text 1 had a factor loading as low as .020, whereas the highest factor loading on text 1 was .624. These factor loadings reveal that the items do not share much common variance after controlling for the general factor. The variance explained by the latent factors is also displayed in Table 4. The general factor explained most of the variance in the items, ranging from 18.884 to 27.646. The variance explained by the latent text factors was small, but reasonable (range from 2.747 to 14.152), whereas the latent inference factors explained little additional variance (range 2.137 to 6.102). This suggests that the construct of inference making is broadly uni-dimensional: although there is some statistical confirmation that inferences that serve different coherence functions are statistically distinguishable (based on the model comparisons), their additional explanatory power is limited. Lastly, the reliabilities of the different latent factors were calculated. The reliability of the general factor was high, ranging from .779 to .845. The latent text factors as well as the latent inference making factors were less reliable (reliability of the text factors: .011 to .748; reliability of the inference making factors: .000 to .339). This shows that the separate types of inference (local and global coherence) could not be measured reliably. These results were replicated when testing an alternative model without a general factor[1].

Test Re-Test Reliability of the Inference Task

We calculated test re-test reliability scores to compare with our MTMM modeling findings. Test re-test reliability was calculated by correlating the local and global inference making composite scores in two consecutive years. The correlations of local and global inference making between consecutive years were moderate to high: local inferences (P to K, r = .54; K to grade 1, r = .38; grade 1 to 2, r = .38; grade 2 to 3, r = .45) and global inferences (P to K, r = .53; K to grade 1, r = .54; grade 1 to 2, r = .51; grade 2 to 3, r = .46). In contrast, the correlations between the total raw scores of the Inference Task for consecutive years were consistently high (P to K, r = .63; K to grade 1, r = .58; grade 1 to 2, r = .56; grade 2 to 3, r = .54). Considering the fact that there is one year in between those measurement occasions, and that inference making is determined by other variables that change with age such as vocabulary, memory, and strategy knowledge, we argue that the Inference Task as a whole is sufficiently reliable, at least for research purposes. The weaker correlations for inference type between years indicates lower reliability. We expand on this point in the Discussion.

Validity of the Inference Task

Since the MTMM modeling and the test re-test reliability analyses did not indicate reliable and separable measures for local and global inferences, we examined the validity of the Inference Task as a whole. The correlations between the sum scores of the Inference Task and the three listening comprehension measures in all grades are presented in Table 2. The correlations between the Inference Task and the three listening comprehension measures were moderate to high in all grades. Of note, one of the listening comprehension assessments (LCM) had sufficient numbers of literal and inferential (not distinguished by inference type) questions to extract composite scores. The correlations between the scores for those conceptually different aspects of text comprehension were of the same magnitude as the correlations between different types of inference on our experimental measure (*r*s = .49 - .69).

Due to the low reliability of some of the listening comprehension measures, the correlations between the composite scores vary to a large extent. Therefore, factor scores were extracted from one-factor models with the three listening comprehension measures for each grade. These models consist of one general listening comprehension factor with three indicators: the listening comprehension measures. These models had zero degrees of freedom, thus all models are just-identified. The correlations between the sum scores of the Inference Task with the listening comprehension factor scores were high in all grades (P: *r* = .76; K: *r* = .73; G1: *r* = .56; G2: *r* = .62; G3: *r* = .58). These high correlations indicate that the Inference Task is a valid measure to assess listening comprehension, in general.

Discussion

We examined the dimensionality of inference making and the validity of an experimental measure used to assess inference making. Our texts were constructed to assess both local and global coherence inferences. We found support for the distinction of these two types of inference; our best fitting model included these inference types as separate factors.

However, this finding was qualified by the fact that the two inference types explained only a small amount of unique variance on the task relative to a factor that represented the text as a whole. In addition, reliability analyses questioned the extent to which the two types of inference, when embedded in the same text, could be measured reliably. Of note, relations between our experimental measure of inference making and several listening comprehension measures confirmed that our inference measure as a valid measure of discourse level listening comprehension.

As a first aim, the dimensionality of inference making in the oral domain was investigated. The best fitting model for each grade (the MTMM model) took into account the fact that our items to assess local and global coherence making were nested within texts, and also included a general inference factor. Thus, statistically, there was evidence that local and global coherence inferences can be distinguished. However, of note, the general inference making factor, which included both types of inference, was more reliable than the separate inference types, and also explained a greater proportion of the variance (on average just over 20 percent of the unique variance). This is in line with previous studies that have investigated the dimensionality of reading comprehension (e.g., Basaraba et al., 2013; Muijselaar et al., 2017). Most of the text factors (i.e., concerning the text to which the items belonged) were not reliable, but these factors explained a reasonable percentage of unique variance, on average just over 6 percent. However, the corresponding factor loadings were very low and sometimes even negative indicating that questions about the same text do not necessarily share common variance. This was confirmed by the fact that these text factors did not in general reach acceptable levels of reliability. Although the local and global inference factors explained on average four percent of the variance, the factor loadings of these factors were low and sometimes negative and the latent local and global inference factors were unreliable. Test re-test reliability analyses for the separate inference factors were moderate. Thus, although models that included both text and inference factors were the best fitting models, we cannot reliably distinguish between texts and different coherence inferences.

One reason for the lack of clear distinction and reliable measurement of text and coherence inference function comes from a consideration of the product of text comprehension. Theories of text comprehension concur that successful understanding results in an integrated and coherent representation of the text's meaning, typically referred to as a mental model or situation model (see McNamara & Magliano, 2009, for a review). To construct this representation, skilled comprehenders encode literal details and, critically, they strive for cohesion and coherence (van den Broek, Risden, & Husebye-Hartman, 1995) by integrating the ideas presented in the text and generating inferences using their background knowledge (Cain & Oakhill, 1999; Elbro & Buch-Iverson, 2013; Graesser et al., 1994). Thus, successful text comprehension relies on memory for explicit facts (literal details) in a text, as well as inferences to establish local and global coherence. Because these two functions of inference making are both deemed necessary for successful reading and listening comprehension (Graesser et al., 1994), they may be hard to distinguish. A related reason for differences in factor loadings for different inference items might be that both local and global coherence inferences can be required to understand fully the same episode in a story. Thus, if a listener or reader builds a particularly detailed and integrated representation of one part of a story, these two aspects of inference making will share common variance, which has nothing to do with the separate inference factors on which these items both load. In addition, the mental model constructed of the text serves as the context in which to interpret subsequent portions of text. As a result, a representation that is not fully coherent may influence the likelihood of further inference making.

An additional different reason for the lack of clear distinction between local and global coherence inferences and also for the lack of homogeneity in their measurement is that within each function of inference, a range of different inference types can be made (Graesser et al., 1994). For example, local coherence inferences are generated to establish reference, e.g., ‘He fetched a glass of orange juice. The drink was very refreshing.’ (Cain & Oakhill, 1999) and also to establish causality ‘Dorothy poured the bucket of water on the bonfire.’ (Singer & Halldorson, 1996) (see Graesser et al., 1994, for additional types). Similarly, global coherence inferences can involve establishing superordinate goals, themes, emotional responses etc. (Graesser et al., 1994). Within a given category of inference, there can also be variation in the distance between the inference and supporting context as shown by Myers, Shinjo, and Duffy’s (1987) study of causal relatedness, which influences memory and comprehension of the text. Because of that variety within each type of necessary coherence inference, strong internal consistency might be hard to find for the inferences targeted in the narrative texts used in this study.

Finally, we should consider how inferences depend on vocabulary and background knowledge, factors that may influence inference making performance but which were not included in our analyses. Text integration to achieve local coherence occurs as skilled comprehenders attempt to integrate the meaning of a just-presented sentence with their current mental model and may often be signaled by category-exemplar pairings as in the ‘orange juice – drink’ example above. Successful integration often relies on vocabulary (or background) knowledge (Cain & Oakhill, 2014; Perfetti, Yang, & Schmalhofer, 2008), and indeed, Perfetti and colleagues have argued that individual differences in the quality of semantic representations underpin ease and accuracy of integration. For example, younger children or those with less rich and precise semantic (or general) knowledge, may not readily activate related concepts as they hear or read a text. Global coherence inferences also draw on vocabulary and general knowledge to make full sense of details that have been left unstated (Graesser et al., 1994), although they are not always signaled by a noun phrase or anaphoric reference (see Cain & Oakhill, 2014, for further discussion). Less-skilled adult comprehenders are less likely to draw on their general knowledge as they read, in order to make global coherence inferences (Long, Oppy, & Seely, 1994), and knowledge accessibility is a factor determining young children's inference making (Barnes, Dennis, & Haefele-Kalvaitis, 1996). However, knowledge also varies across individuals and it is well established that both adult and child experts in a given subject area can 'compensate' for poor general text processing skills (Schneider, Körkel, & Weinert, 1989; Spilich, Vesonder, Chiesi, & Voss, 1979). Thus, an individual's knowledge may be critical for generating individual inferences (see Compton, Miller, Elleman, & Steacy, 2014, for a broader discussion of the role of knowledge in reading development). There is some evidence, that vocabulary predicts the ability to generate global coherence inferences more strongly than local cohesion (the term used by the authors) inference (Cain & Oakhill, 2014). Future research needs to examine the extent to which individual differences in vocabulary and/or general knowledge determine the ability to generate both local and global coherence inferences.

In sum, the local and global inference factors only explained a small (but significant) part of the variance of inference making, and these separate factors could not be measured reliably. One explanation for the finding that local and global inferences could not be distinguished is that both are both necessary for comprehension of the same text. Therefore, when tested for the same text, they are likely to be highly related. In addition, a range of different inference types are associated with local and to global coherence, and both draw on external knowledge. Thus, the very nature of text comprehension and inferences may make it difficult to design an assessment in which these two aspects of inference making can be statistically distinguished.

The second aim of this study was to examine the validity of the Inference Task in relation to discourse-level listening comprehension. The correlations between performance on the Inference Task and the listening comprehension factor scores were high, and the correlations with the composite scores of the different listening comprehension measures ranged from moderate to high. Inference making is essential to the construction of a mental model of a text, but other skills are also important such as recall of explicit details in the text is also essential (Cain & Oakhill, 2014), and knowledge and use of story knowledge and comprehension monitoring explain unique variance (Kim, 2016; Oakhill & Cain, 2012). For that reason, these moderate to high correlations are sufficient to indicate that the Inference Task is a valid measure of this particular aspect of discourse listening comprehension.

Our findings must be interpreted in the context of some constraints of the present study. A first limitation concerns the low reliability of some of the listening comprehension measures. This was however taken into account by using factor scores in addition to composites for the validity analyses. Further, the small number of items at each grade level (8 local and 8 global inference questions) might result in unreliable specific factors. However, a previous study (Muijselaar et al., 2017) in which a bigger set of items was analyzed to examine the dimensionality of reading comprehension resulted in the same conclusion, namely that specific aspects of reading comprehension (question types) could not reliably be distinguished. In addition, we were able to investigate whether the Inference Task is a valid measure in relation to listening comprehension. Future studies should investigate whether this measure is also a valid measure of inference making, by correlating it to different listening comprehension and inference making measures. Lastly, in this study, very young children were sampled compared to other studies of inference making in children, which have typically looked at 7 through 11 years of age and considered reading, rather than listening, comprehension (e.g., Cain & Oakhill, 2014). Subtypes of necessary inference making, and indeed other measures of listening comprehension such as literal comprehension, may be more evident in older readers when they have had greater experience comprehending more complex texts that draw on a range of processes and knowledge bases (e.g., Goldman, 2012).

An implication of the present study concerns the training of inference making. Our study showed that inferences that serve different functions are not necessarily homogeneous. This implies that the training of specific *functions* of inferences may not affect performance on specific question types, but will instead influence performance on the entire listening comprehension test. Indeed, training in more general comprehension skills, such as inference and narrative structure does lead to improved performance on more global measures of comprehension such as standardised assessment (Clarke, Snowling, Truelove, & Hulme, 2010; Oakhill & Sullivan, 2016). This is supported by a recent meta-analysis of inference interventions, which additionally found that the benefits for poorer readers can extend to performance on literal questions (Elleman, 2017). Given the wide range of inference types, it may be prudent to determine whether training should focus on specific inference types, specific inference skills, or a more general standard of coherence (van den Broek et al., 1995). As noted, previous research with older children has shown differential prediction by vocabulary of these two functions of inference (Cain & Oakhill, 2014). Thus, future research should investigate whether inferences that serve different functions (e.g., local vs global coherence) and/or different inference types (e.g., referential, causal antecedent, superordinate goals, instantiation of nouns and verbs, etc.) can be more reliably distinguished in older children.

In sum, local and global coherence inferences could not be reliably distinguished. Critically, our experimenter-developed Inference Task is a valid measure to assess discourse listening comprehension. This implies that the Inference Task could be used as a measure for general listening comprehension, but that the distinction of specific functions of necessary inference may not be required.

Footnotes.

1. We tested an alternative model to replicate the results of Model 3. In this alternative model items load on one of the two correlated text factors and one of the correlated inference factors. The variance explained by the text factors ranged from 5.974% to 23.390%; the variance explained by the inference factors was between 4.837% and 11.307%. The reliability of the text factors was .364 to .869; the reliability of the inference factors ranged from .068 to .686. The inference factors were sufficiently reliable in grades 1 and 2 (ranging from .619 to .686), however, this sufficient reliability could not be replicated in the other grades. This higher reliability in grades 1 and 2 can also be explained by the fact that the text factors are very unreliable. Thus, despite that removing the general inference making factor results in more variance explained by the inference making factors, and higher reliabilities, we should still conclude that local and global inferences could not be distinguished reliably in all grades.

References

Barnes, M. A., Dennis, M., & Haefele-Kalvaitis, J. (1996). The effects of knowledge availability and knowledge accessibility on coherence and elaborative inferencing in children from six to fifteen years of age. *Journal of Experimental Child Psychology, 61*, 216-241. doi: 10.1006/jecp.1996.0015

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing*, *26*, 349-379. doi:10.1007/s11145-012-9372-9

Brown, R. L. (1989). Using covariance modeling for estimating reliability on scales with ordered polytomous variables. *Educational and Psychological Measurement*, *49*, 385-398. doi:10.1177/0013164489492011

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 136-162). Newbury Park, CA: Sage.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*, 221-248. doi:10.1037/a0023350

Cain, K., & Oakhill, J. V. (1999). Inference making and its relation to comprehension failure. *Reading and Writing. An Interdisciplinary Journal*, *11*, 489-503. doi:10.1023/A:1008084120205

Cain, K., & Oakhill, J. (2014). Reading comprehension and vocabulary: Is vocabulary more important for some aspects of comprehension? *L'Année Psychologique, 114*, 647-662.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, *96*, 31-42. doi:10.1037/0022-0663.96.1.31

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence*, (pp. 5-22). Oxford, United Kingdom: Pergamon.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14,* 464–504. doi:10.1080/10705510701301834

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*, 189-225. doi:10.1207/s15327906mbr4102_5

Chrysochoou, E., Bablekou, Z., & Tsigilis, N. (2011). Working memory contributions to reading comprehension components in middle childhood children. *American Journal of Psychology, 124*, 275-289.

Compton, D. L., Miller, A. C., Elleman, A. M., & Steacy, L. M. (2014). Have we forsaken reading theory in the name of "quick fix" interventions for children with reading disability? *Scientific Studies of Reading*, *18*, 55-73. doi: 10.1080/10888438.2013.836200

Currie, N. K., & Cain, K. (2015). Children's inference generation: the role of vocabulary and working memory. *Journal of Experimental Child Psychology, 137*, 57-75. doi:10.1016/j.jecp.2015.03.005

Davis, F. B. (1944). Fundamental factors of comprehension in reading. *Psychometrika*, *9*, 185-197. doi:10.1007/BF02288722

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, *13*, 230-253. doi:10.1037/a0013219

Elleman, A. M. (in press). Examining the impact of inference instruction on the literal and inferential comprehension of skilled and less skilled readers: A meta-analytic review. *Journal of Educational Psychology.* doi: 10.1037/edu0000180

Freed, J. & Cain, K. (2017). Assessing school-aged children's inference making: The effect of test story format in listening comprehension. *International Journal of Language and Communications Disorders 52,* 95-105. doi: 10.1111/1460-6984.12260

Gibbons, R. D., Darrell Bock, R., Hedeker, D. R., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . . Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*, 4-19. doi:10.1177/0146621606289485

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423-436. doi:10.1007/BF02295430

Gillam, R. B., & Pearson, N. A. (2004). Test of Narrative Language–Receptive. Austin, TX: Pro-Ed.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371-395. doi: 10.1037/0033-295X.101.3.371

Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*, 179-203. doi:10.1016/0160-2896(84)90008-4

Gustafsson, J.-E. (2002). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 73-95). Mahwah, NJ: Erlbaum.

Gustafsson, J.-E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97-121). Washington, DC: American Psychological Association.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. doi:10.1080/10705519909540118

Kim, Y. S. G. (2016). Direct and mediated effects of language and cognitive skills on comprehension of oral narrative texts (listening comprehension) for children. *Journal of Experimental Child Psychology*, *141*, 101-120. doi:10.1016/j.jecp.2015.08.003

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363-394.

Kline, R. B. (2011). *Principles and practice of structural equation modeling (3rd ed.)*. New York, NY: Guilford Press.

Language and Reading Research Consortium (LARRC). The dimensionality of language ability in young children. *Child Development, 86,* 1948-1965. doi: 10.1111/cdev.12450

Language and Reading Research Consortium (LARRC), Farquharson, K., & Murphy, K. A. (2016). Ten steps to conducting a large, multi-site, longitudinal investigation of language and reading in young children. *Frontiers in Psychology*, *7*, 1-16. doi:10.3389/fpsyg.2016.00419

Leslie, L., & Caldwell, J. S. (2010). Qualitative Reading Inventory (5th ed.). Boston, MA: Pearson.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: The Guildford Press.

Long, D. L., & Chong, J. L. (2001). Comprehension skill and global coherence: A paradoxical picture of poor comprehenders' abilities. *Journal of Experimental Psychology*, *27*, 1424-1429. doi:10.1037/0278-7393.27.6.1424

Maruyama, G. M. (1998). *Basics of structural equation modeling.* Thousand Oaks, CA: Sage Publications.

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology*, *4*, 1-13. doi:10.3389/fpsyg.2013.00169

Muijselaar, M. M. L., Swart, N. M., Steenbeek-Planting, E. G., Droop, M., Verhoeven, L., & de Jong, P. F. (2017). The dimensions of reading comprehension in Dutch children: Is differentiation by text and question type necessary? *Journal of Educational Psychology*, *109*, 70-83. doi:10.1037/edu0000120

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide*. *Seventh edition*. Los Angeles, CA: Muthén & Muthén.

Myers, J. L., Shinjo, M., & Duffy, S. A. (1987). Degree of causal relatedness and memory. *Journal of Memory and Language*, *26*, 453-465.

Oakhill, J., & Cain, K. (2012). The precursors of reading comprehension and word reading in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading, 16*, 91-121, doi: 10.1080/10888438.2010.529219

Oakhill, J., & Sullivan, S. (2016). The relation between understanding text connectives and reading comprehension in novice readers. Paper presented at the annual conference of the Society for the Scientific Study of Reading, Porto, Portugal.

Paris, A. H., & Paris, S. G. (2003). Assessing narrative comprehension in young children. *Reading Research Quarterly*, *38*, 36-76. doi: 10.1598/RRQ.38.1.3

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53-61. doi:10.1007/bf02289209

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals* (4th ed.). Bloomington, MN: Pearson.

Silva, M., & Cain, K. (2015). The relations between lower and higher level comprehension skills and their role in prediction of early reading comprehension. *Journal of Educational Psychology*, *107*, 321. doi: 10.1037/a0037769

Spearritt, D. (1972). Identification of sub-skills of reading comprehension by maximum likelihood factor analysis. *ETS Research Bulletin Series*, *1972*, i-24. doi:10.1002/j.2333-8504.1972.tb00192.x

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics (6th ed.).* Boston: Allyn and Bacon.

Thorndike, R. L. (1973). *Reading Comprehension Education in Fifteen Countries: An empirical study. International Studies in Evaluation III.* Uppsala, Sweden: Almqvist and Wiksell.

Undheim, J. O., & Gustafsson, J. (1987). The hierarchical organization of cognitive abilities: Restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research*, *22*, 149-171. doi:10.1207/s15327906mbr2202_2

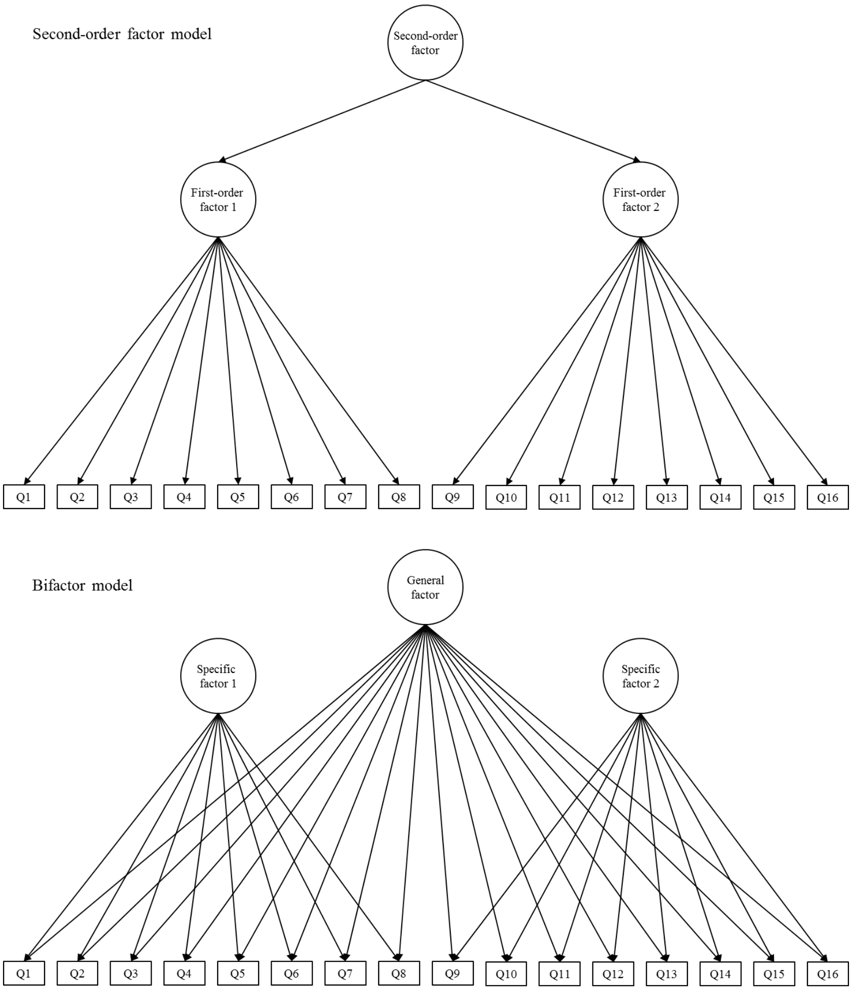

Figure 1. A second-order factor model and a bifactor model.

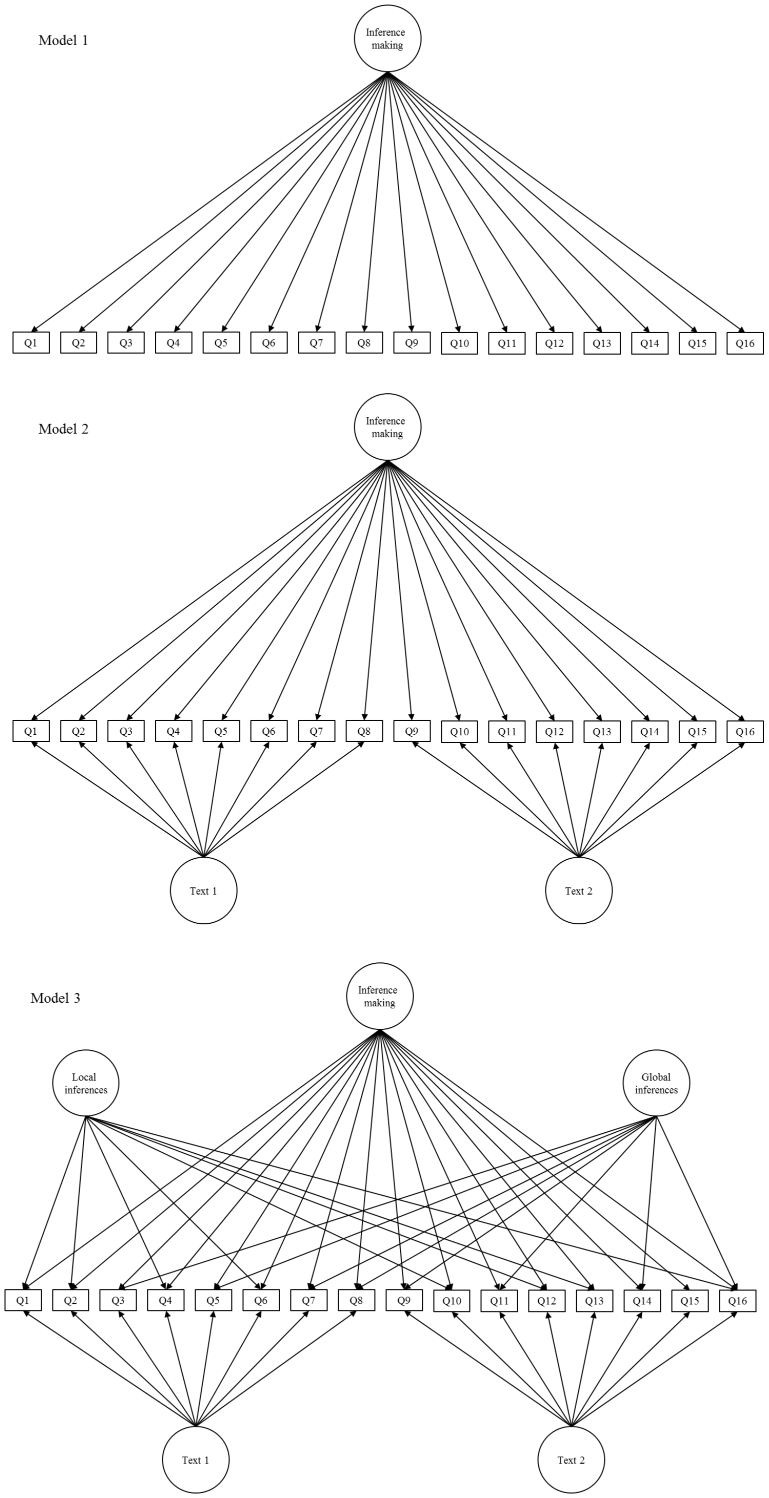

Figure 2. The one-factor model (Model 1), bifactor model (Model 2), and MTMM model (Model 3).

Table 1

Descriptive Statistics by Grade for all Observed Variables

	P			K			G1			G2			G3		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Inference Task	327	0.83	0.42	469	1.11	0.37	547	1.18	0.36	654	1.19	0.38	664	1.53	0.31
Local	345	0.88	0.49	474	1.05	0.44	553	1.15	0.41	658	1.17	0.43	672	1.42	0.41
Global	292	0.76	0.40	467	1.17	0.40	546	1.22	0.42	650	1.22	0.42	660	1.63	0.31
LCM	382	7.00	3.28	465	9.98	2.43	557	11.85	2.42	654	19.53	4.31	675	20.92	4.99
Implicit	382	2.24	1.38	473	3.25	1.25	563	4.04	1.35	663	9.03	2.98	679	9.43	2.81
Explicit	382	4.76	2.24	473	6.57	1.83	563	7.70	1.69	663	10.27	2.26	679	11.39	2.79
USP	376	5.91	2.28	481	6.41	1.83	563	6.78	1.29	669	6.60	1.67	669	7.34	1.88
TNL	370	15.65	6.98	473	23.05	5.47	551	27.15	4.15	656	29.69	3.94	674	31.37	3.45

Note. LCM: Listening Comprehension Measure; USP: CELF 4, Subtest Understanding Spoken Paragraphs; TNL: Test of Narrative Language – Receptive.

Table 2

Pearson Correlations between all Observed Variables

P	1	1a	1b	2	3	4
1. Inference Task	1					
1a. Local	.94	1				
1b. Global	.89	.67	1			
2. LCM	.69	.64	.65	1		
3. USP	.61	.57	.53	.63	1	
4. TNL	.72	.68	.63	.76	.62	1

K	1	1a	1b	2	3	4
1. Inference Task	1					
1a. Local	.91	1				
1b. Global	.88	.61	1			
2. LCM	.64	.56	.57	1		
3. USP	.45	.39	.42	.46	1	
4. TNL	.67	.60	.59	.64	.41	1

G1	1	1a	1b	2	3	4
1. Inference Task	1					
1a. Local	.86	1				
1b. Global	.87	.49	1			
2. LCM	.48	.36	.46	1		
3. USP	.37	.30	.32	.29	1	
4. TNL	.46	.35	.43	.53	.31	1

G2	1	1a	1b	2	3	4
1. Inference Task	1					

1a. Local	.89	1				
1b. Global	.88	.57	1			
2. LCM	.59	.50	.55	1		
3. USP	.42	.38	.35	.39	1	
4. TNL	.45	.39	.38	.53	.38	1

G3	1	1a	1b	2	3	4
1. Inference Task	1					
1a. Local	.90	1				
1b. Global	.82	.50	1			
2. LCM	.54	.47	.47	1		
3. USP	.42	.37	.37	.46	1	
4. TNL	.40	.37	.31	.50	.38	1

Note. All correlations are significant (*p* < .01)

LCM: Listening Comprehension Measure; USP: CELF 4, Subtest Understanding Spoken Paragraphs; TNL: Test of Narrative Language – Receptive.

Table 3

Fit Indices for the Confirmatory Factor Models of the Dimensionality of Inference Making in all Grades

Grade	Model		χ^2	*df*	χ^2/df	RMSEA	90% CI	CFI	Model comparison	$\Delta\ \chi^{2\ \#}$	$\Delta\ df$
P	1	1 factor	191.728**	104	1.844	.047	.036 - .057	.944	-	-	-
	2	1 factor + 2 texts	151.652**	89	1.704	.043	.031 - .054	.960	1 vs 2	41.450**	15
	3	1 factor + 2 texts + 2 inferences	107.386**	73	1.471	.035	.019 - .048	.978	2 vs 3	45.255**	16
K	1	1 factor	190.098**	104	1.828	.041	.032 - .050	.922	-	-	-
	2	1 factor + 2 texts	106.914	88	1.215	.021	.000 - .034	.983	1 vs 2	76.475**	16
	3	1 factor + 2 texts + 2 inferences	76.318	72	1.060	.011	.000 - .029	.996	2 vs 3	31.340*	16
G1	1	1 factor	221.314**	104	2.218	.044	.036 - .052	.892	-	-	-
	2	1 factor + 2 texts	125.895**	88	1.431	.027	.015 - .038	.965	1 vs 2	85.963**	16
	3	1 factor + 2 texts + 2 inferences	95.902*	72	1.332	.024	.008 - .036	.978	2 vs 3	29.350*	16
G2	1	1 factor	219.209**	104	2.108	.040	.033 - .048	.933	-	-	-
	2	1 factor + 2 texts	111.192*	88	1.264	.020	.002 -.030	.987	1 vs 2	90.798**	16
	3	1 factor + 2 texts + 2 inferences	83.203	72	1.156	.015	.000 - .028	.993	2 vs 3	27.051*	16

G3	1	1 factor	308.155**	105	2.935	.053	.046 - .060	.915	-	-	-
	2	1 factor + 2 texts	141.010**	89	1.584	.029	.020 - .038	.978	1 vs 2	126.917**	16
	3	1 factor + 2 texts + 2 inferences	78.826	73	1.080	.009	.000 - .024	.998	2 vs 3	65.197**	16

#Corrected chi-square difference test (Satorra-Bentler correction).

**p* < .05; ***p* < .01.

Table 4

Factor Loadings, Variance Explained, and Reliabilities of the Latent Factors of the Final Model for all Grades

Grade		*k*	Median λ	Min λ	Max λ	R^2 (%)	*ρ*
P	General factor	16	.559	.033	.728	27.646	.845
	Text 1	8	.220	.020	.624	4.670	.336
	Text 2	8	.270	.000	.545	5.270	.416
	Local inferences	8	.029	-.358	.296	2.351	.000
	Global inferences	8	.084	-.036	.733	5.689	.294
K	General factor	16	.438	.090	.603	19.635	.780
	Text 1	8	.256	.017	.687	6.991	.450
	Text 2	8	.082	-.338	.598	4.969	.118
	Local inferences	8	.089	-.272	.543	2.648	.073
	Global inferences	8	.025	-.539	.361	3.169	.000
G1	General factor	16	.439	.251	.615	18.884	.779
	Text 1	8	.001	-.441	.524	4.743	.011
	Text 2	8	.236	.073	.674	5.164	.402

	Local inferences	8	.106	-.062	.661	6.102	.339
	Global inferences	8	.058	-.271	.577	4.118	.101
G2	General factor	16	.463	.225	.704	21.763	.805
	Text 1	8	.387	.219	.523	7.851	.583
	Text 2	8	.139	-.197	.393	2.747	.113
	Local inferences	8	.181	-.014	.536	3.890	.262
	Global inferences	8	.015	-.237	.445	2.137	.011
G3	General factor	16	.471	.300	.596	23.100	.822
	Text 1	8	.118	-.505	.388	3.981	.060
	Text 2	8	.497	.343	.719	14.152	.748
	Local inferences	8	-.003	-.299	.780	4.790	.025
	Global inferences	8	.170	-.093	.731	5.086	.269

Appendix A

Stories and questions (* = local coherence inference; other questions tap global coherence inference) by grade

BIRTHDAY (P)

Today was Grandma's birthday. The family was getting ready for the party. Dad and Josh were putting up the party tent in the back lawn. Mom told them to put on some sunscreen, so that they didn't burn.

Mom drove over to pick up Grandma, who lived an hour away. Mom told Linzie to keep an eye on the cake in the oven and to make some fruit punch. Linzie was slicing oranges when the knife slipped. Her finger was bleeding but she couldn't find any bandages! Luckily, Brenda, their next-door neighbor, had some.

Back in the house, the kitchen was filled with smoke. Linzie looked in the oven. Oh dear! Mom would be mad. Then, Linzie had an idea. She drove to the grocery store.

When she got back home, her aunts, uncles, and cousins were all waiting quietly in the party tent. Linzie put her purchase at the center of the dessert table. A few minutes later, Mom walked into the party tent with Grandma. Everything looked perfect. Grandma was amazed that all of her family was there. "What a wonderful surprise." she said.

Questions (examples of full scoring range provided for first two questions)

1. What were the family getting ready for? *

Answer: Grandma's (birthday) party (2 points); a party (1 point); to go out (0 points)

2. What was the weather like?

Answer: (hot and) sunny (2 points); warm (1 point); rainy (0 points)

3. Why did Linzie need some bandages? *

Answer: cut her finger (on knife)

4. Where did Linzie get the bandages?

Answer: next door, from Brenda, from neighbor etc.

5. Why would Mom be mad?

Answer: cake was burnt, ruined

6. Why did Linzie drive to the grocery store? *

Answer: to get/buy a (new) cake/she needed a new cake

7. Where was the dessert table? *

Answer: in the party tent

8. Why was Grandma surprised?

Answer: didn't know they were having a party for her

A NEW PET (P & K)

Tim had a new pet called Sparky. Sparky was soft, furry, and very playful. At first, Sparky slept indoors in a cardboard box with a nice soft blanket. Sparky soon grew very big. Tim decided to build a kennel and a tall wooden fence around the back yard.

Tim went to the store. He already had a hammer and a saw, but he needed some wood and some nails. Tim built the kennel first. His friend Jack helped him to build the fence. Jack held the wood and Tim banged in the nails. The fence was soon finished. Even though Tim's thumb was bruised and sore, he was smiling. He put the hammer that had caused the pain away in his toolbox. He was very pleased with his hard work.

That evening, Tim moved Sparky into his new home. But, Sparky did not like his new home. His old cardboard box was still indoors and Sparky missed his nice soft blanket.

Questions

1. What sort of animal was Sparky?

Answer: dog

2. What did Tim buy at the store? *

Answer: wood and nails

3. Who put up the fence? *

Answer: Tim & Jack, Tim & his friend, the man & his friend

4. Why did Tim need a tall fence?

Answer: because Sparky could jump/ so Sparky didn’t run away

5. Why did Tim have a sore thumb?

Answer: banged/hit his thumb with hammer etc.

6. Where was Sparky’s kennel? *

Answer: in yard, outside in back yard

7. Why did Sparky no longer sleep in the cardboard box?

Answer: he was too big, he had grown too big, outgrown it

8. Where was Sparky’s blanket? *

Answer: (still) in his box, in the house

THE GAME (K & G1)

Today was the last game of the season. There was only a minute left and the score was tied. Jake ran towards the goal and kicked the ball passed the goalie. He had scored a goal. The crowd cheered. Jake’s team had won.

After the game, Jake got his shampoo and towel and took a shower. That felt good. He put the towel in his backpack and his other things away in his locker, and set off for home. It was usually just a 10-minute ride, but Jake was feeling tired. He pedaled slowly and the trip took more than 20 minutes.

Jake was starving when he got home, so he searched the cupboards. He found what he was looking for on the top shelf. The cookie jar was full because Mom had been shopping. Mom cooked Jake his favorite food for dinner that evening, but Jake wasn't hungry. He couldn’t finish his burger and fries.

Questions

1. What sport was Jake playing?

Answer: soccer

2. By how many goals did Jake’s side win? *

Answer: 1

3. Where did Jake put his shampoo after he had showered? *

Answer: locker

4. How did Jake get home?

Answer: bicycle, bike, he cycled

5. Why was Jake tired?

Answer: because he had played soccer

6. Where was the cookie jar? *

Answer: (top shelf of) cupboard

7. Why wasn’t Jake hungry at dinner time? *

Answer: eaten too many cookies, full of cookies etc.

8. What was Jake’s favorite food?

Answer: hamburger and/or fries

STEVE AND DAD’S DAY OUT (G1 & G2)

Dad parked the car. He had picked a nice shady spot by some trees. Dad and Steve opened the trunk. They got out their things. First, they put on their boots. Then they checked their backpacks. It was a hot day and they each had plenty of water.

Dad and Steve followed the trail up through the forest. After the trees, the trail got steep and rocky. They were pleased they had lots of water. Steve was very fit and he was soon out of sight. The rocks on the path were loose and Steve slipped. He cried out. Dad came running up behind.

Steve was not hurt. They had a short break and walked on. Dad and Steve were tired when they got to the top, but it was worth the effort. They could see for miles.

Questions

1. Where did Dad park the car? *

Answer: in the shade; by some trees

2. What was in the trunk of the car? *

Answer: any of boots, backpacks, or water

3. Was the rocky path before or after the trees? *

Answer: after

4. Why were they pleased that they had brought lots of water?

Answer: it was hot day; hiking was thirsty work

5. Why was Steve soon out of sight? *

Answer: faster; quicker; fitter than dad etc

6. Why did Steve slip?

Answer: loose rocks, lost his footing, tripped on rocks

7. How did Dad know that something was wrong?

Answer: heard Steve cry out, Steve yelled, etc.

8. What were they doing?

Answer: hiking; climbing a mountain

ANN AND DAVID (G2 & G3)

It was getting dark when Ann and David walked into the park. They had been looking forward to this all day. They were going to have fun. They went on lots of different rides. Ann liked the Rocket Ride best, but David thought the Roller Coaster was scarier. They got hungry so they walked back to the entrance and joined the line for hotdogs.

Ann and David had now spent the last of their money. They walked over to the lake. A big crowd was waiting for the display to begin. This was the big event. First, there was music. Then the sky lit up with flashes and flares and bangs. And then, sadly, it was over.

Ann and David had no money for a bus. They had to walk home. On their way back, the weather changed. They were very wet by the time they got home.

Questions

1. What time of day was it?

Answer: evening, night-time, dusk

2. Where did Ann and David go?

Answer: funfair, amusement park

3. Where was the hot dog stand? *

Answer: near entrance to park, by the entrance

4. Where was the crowd? *

Answer: by the lake

5. What display was everyone waiting for?

Answer: fireworks

6. How did the display begin? *

Answer: music

7. Why did they have no money for a bus? *

Answer: spent last on hot dogs, spent on rides and hot dogs

8. Why did they get wet on the way home?

Answer: it rained, there was a rainstorm, etc.

A FAMILY DAY OUT (G3)

Billy, Susie, and their Mom had gone out for the day. Billy spent the morning building a sandcastle near the water. Mom sat on their large beach towel and read a book. Susie wanted to go for a swim. She put her feet in the sea but the water felt too cold. Susie went and sat down next to Mom, instead. Mom had packed a big bag full of books and games. Susie found her story book and started to read.

Suddenly, they heard Billy crying. Mom and Susie looked over. Billy's day was ruined. All his hard work had been washed away by a wave.

Mom ran over to Billy, but he did not stop crying. Susie remembered the ice cream stand. She ran over to the parking lot and bought Billy his favorite flavor. Billy smiled when he saw his sister walking over. Soon the tears stopped.

Questions

1. Where were Billy and his family?

Answer: beach, seaside, coast

2. Why did Susie not swim in the sea?

Answer: Because the water was too cold

3. Where did Susie find her book? *

Answer: In the bag (of books and toys)

4. What did Susie sit on? *

Answer: beach towel.

5. Why was Billy crying?

Answer: A wave had broken his sandcastle, sandcastle washed away, etc.

6. Why did mom run over to Billy?

Answer: To comfort him; see what was wrong

7. Where was the ice cream stand? *

Answer: by the parking lot

8. Why did Billy smile when he saw his sister? *

Answer: His sister was carrying his favorite (flavor) ice cream etc.