Hammill Institute on Disabilities

Assessment for Effective Intervention
2021, Vol. 47(1) 47–52

DOI: 10.1177/1534508420937781
aei.sagepub.com
SAGE

Development and Validation of the Minnesota Inference Assessment

Panayiota Kendeou, PhD[1], Kristen L. McMaster, PhD[1], Reese Butterfuss, MA[1], Jasmine Kim, MA[1], Susan Slater, MA[1], and Okan Bulut, PhD[2]

Abstract

The overall aim of the current investigation was to develop and validate the initial version of the Minnesota Inference Assessment (MIA). MIA is a web-based measure of inference processes in Grades K–2. MIA leverages the affordances of different media to evaluate inference processes in a nonreading context, using age-appropriate fiction and nonfiction videos coupled with questioning. We evaluated MIA's technical adequacy in a proof-of-concept study. Taken together, the results support the interpretation that MIA shows promise as a valid and reliable measure of inferencing in a nonreading context for students in Grades K–2. Future directions involve further development of multiple, parallel forms that can be used for progress monitoring in K–2.

Keywords

inference, assessment, comprehension

Theories of comprehension specify how comprehension emerges during moment-by-moment processing and assume that it is based on a mental representation that is constructed over the course of reading (e.g., Kintsch, 1998). Most theories also emphasize inferences as *core processes* because they are *necessary* for the construction of a coherent situation model, which is the *product* of reading comprehension (McNamara & Magliano, 2009). Two types of inferences are necessary in this context (Graesser, 2015; McNamara, 2004): *Bridging inferences*, which involve connections between ideas within the text, and *elaborative inferences*, which involve connections between information in the text and prior knowledge (Barth et al., 2015).

To understand how inference processes break down for readers who fail to create a coherent mental representation of a text, it is essential to evaluate the construction of this representation *as it occurs moment-by-moment* (Kendeou, 2015; McMaster et al., 2012). However, comprehension is typically assessed *after* reading a text, providing very little insight into the inferential processes *during* reading and why these processes may succeed or fail. As we elaborate next, there are readers who struggle with precisely these processes, but existing measures that examine individual differences in comprehension processes are limited in the information they provide.

Existing Measures of Inference Processes

There are very few measures of inference processes in the context of reading comprehension. One such measure is the Multiple-Choice Online Causal Comprehension Assessment (MOCCA; Biancarosa et al., 2019; Carlson et al., 2014). MOCCA evaluates the processes by which students in Grades 3 to 5 generate causal inferences. For this measure, students read short texts in which one sentence is omitted. Students are presented three sentences and choose which one best fills in the omitted line from each text. The two incorrect responses represent processes on which poor comprehenders have been shown to rely; thus, patterns in response selections provide diagnostic information about the processes students engage in. Carlson et al. reported correlations between MOCCA scores and several reading-related measures (e.g., DIBELS Oral Reading Fluency [ORF]; Woodcock–Johnson [WJ]-III word ID; Curriculum-Based Measurement [CBM] Maze) between $r = -.37$ to .75, and generally acceptable internal consistency (i.e., $\alpha > .60$). A second measure is the Bridging Inferences Test (Bridge-IT; Barth et al., 2015). Bridge-IT was designed to evaluate inferential processes for students in Grades 6 to 12 using an inconsistency paradigm. Students read a set of sentences and judge whether a continuation sentence is consistent with previously read information. Accuracy and response times assess inferential processes. Barth et al. showed that the Bridge-IT accounted

[1]University of Minnesota, Minneapolis, USA
[2]University of Alberta, Edmonton, Canada

Corresponding Author:
Panayiota Kendeou, Department of Educational Psychology, University of Minnesota, 56 East River Road, Minneapolis, MN 55455, USA.
Email: kend0040@umn.edu

for unique variance in reading comprehension (i.e., Gates-MacGinitie Reading Test [GMRT]; MacGinitie & MacGinitie, 1989), with scores also showing good internal consistency (i.e., $\alpha > .80$). Performance on both of these measures, though, also depends on students' decoding ability and, thus, may not provide an accurate index of students' inference skills.

To remove reliance on decoding, measures that evaluate inference processes in nonreading contexts have also been developed. One such measure is the Learning and Reading Research Consortium (LARRC) Inference Task (LARRC, 2015) for Grades preK–4. The test consists of two stories at each grade level, each followed by eight open-ended questions that assess students' ability to generate local and global coherence inferences. The stories and questions were based on the work of Cain and Oakhill (1999). LARRC and Muijselaar (2018) reported correlations between $r = .37$ and .45 with measures of listening comprehension (Clinical Evaluation of Language Fundamentals: Fourth Edition—Understanding Spoken Paragraphs [CELF-4-USP]; Semel et al., 2003) in K–3, and acceptable internal consistency (i.e., $\alpha > .64$). Another measure is the Know-IT (Barnes et al., 1996) for students ages 6 to 15 years. Students are first taught a series of facts about a fictional world called "Gan" and are later tested on those facts. Then 10 episodes/paragraphs of a text about Gan are read aloud to the student followed by four open-ended questions (elaborative inference, bridging inference, literal question, and a simile question) that the student answers orally. Even though performance on both of these measures does not rely on decoding, they must be administered and manually scored in a time-intensive one-on-one context, limiting feasibility for classroom use at scale. Also, both measures have only one form, limiting use to evaluate the effects of instruction or an intervention (e.g., in a pre–post design).

Given the aforementioned limitations of existing measures, new assessments of inferencing are needed that (a) gauge moment-by-moment inferential processes, (b) allow assessment without reliance on decoding, (c) show feasibility for easy and efficient classroom administration, and (d) offer multiple equivalent forms to monitor student progress as well as evaluate the effects of instruction. To address these needs, we developed the Minnesota Inference Assessment (MIA), a web-based measure of inferencing for students in Grades K–2. MIA has a strong theoretical basis. It draws on the Inferential Language Comprehension (iLC) framework (Kendeou et al., 2020), which was proposed to guide the use of visual narratives to teach and assess inferencing skills in educational settings.

The iLC Framework

The iLC framework (Kendeou et al., 2020) proposes that a *general* inferencing skill underlies successful language comprehension and can transfer across contexts and media. Across different media, information that supports the construction of a coherent mental representation can be reactivated using targeted questioning (Graesser & Franklin, 1990). This information can be further integrated through the use of scaffolding and specific feedback. There is empirical evidence for inferencing as a general skill that transfers across media (e.g., Kendeou, 2015). For example, children ages 4 through 8 years generated both bridging and elaborative inferences when asked questions after aural, televised, and written stories (Kendeou et al., 2008). Regardless of media, the inferences that these children generated predicted their overall reading comprehension performance longitudinally. One explanation for transfer of inferencing skill is that learners engage in the same cognitive processes to construct a mental representation of the information across media (Kintsch, 1998; Magliano et al., 2007). Another explanation is that the same underlying text factors (e.g., causal connections, explicated goals, event boundaries) predict comprehension of aural, televised, and written (e.g., Magliano et al., 2012).

In short, the iLC framework proposes that inferencing is a general skill that transfers across media and can be assessed in both reading and nonreading contexts. Consistent with this idea, MIA leverages the processing similarities between reading and nonreading contexts to measure inferencing using videos.

The Present Study

The overall goal of the current investigation was to develop and validate the *initial version* of MIA, a web-based measure of inference processes that does not rely on decoding, for students in K–2. Evidence for technical quality focused specifically on validity, reliability/precision, and intended use of scores (AERA et al., 2014). With respect to validity, we examined the extent to which the difficulty level of MIA items aligned with students' ability levels, thereby indicating that MIA is suitable for its target population. We also examined the extent to which a unidimensional structure fits the data using the Rasch Model (Rasch, 1960) and confirmatory factor analysis (CFA). In addition, we examined criterion-related validity based on correlations with general measures of language and reading comprehension that involved some level of inferencing. With respect to reliability/precision, we examined the internal consistency of MIA using multiple indices (i.e., internal consistency coefficients and person separation index from the Rasch Model). Finally, with respect to the intended use of scores, we explored the adequacy of four parallel, equivalent forms from MIA (created using automated test assembly [ATA] procedures) in evaluating the effects of instruction.

Method

Participants

The current dataset was drawn from a proof-of-concept study examining the efficacy of an inferencing intervention for students in Grades K–2. The current sample consisted of 272 students: Kindergarten (n = 146, 60 female), Grade 1 (n = 102, 52 female), and Grade 2 (n = 24, 11 female) from a Midwestern school district. The sample was racially and ethnically diverse—Kindergarten: 38.4% White, 37% Hispanic, 13% African American, 7.5% Asian/Pacific Islander, and 2.1% American Indian; Grade 1: 15.7% White, 46.1% Hispanic, 23.5% African American, 12.7% Asian/Pacific Islander, and 2% American Indian; Grade 2: 8.3% White, 62.5% Hispanic, and 29.2% African American. The sample was also economically diverse: 59% of Kindergarten students, 74.5% of Grade 1 students, and 87% of Grade 2 students qualified for free or reduced-price lunch.

Measures

MIA. MIA consists of four modules. Each module includes a 5-min video (one fiction, one nonfiction) and 16 inferential multiple-choice questions, each with one correct answer and three meaningful distractors (i.e., incorrect answer choices). Fiction videos were adapted from *Blinky Bill* cartoon episodes (*Apple Thieves* and *Granny's Glasses*). Nonfiction videos were adapted from animal documentaries (*Cephalopods* and *Eagles*). Students can complete each module in approximately 20 min. Questions are administered aurally via an animated pedagogical agent as shown in Figure S1 (see online supplemental materials). The questions interrupt the video to prompt inferences at the point in the video when those inferences are necessary for comprehension (i.e., *online* inferencing). The current version includes two types of inferential questions that are required for comprehension, bridging and elaborative. The process for video selection, editing, and question writing is described in Figure S2 (see Online Supplemental materials).

Language comprehension. The CELF Fifth Edition–Understanding Spoken Paragraphs (CELF-5-USP) subtest (Wiig et al., 2003) was used as a criterion measure. The CELF-5-USP is an individually administered, nationally normed assessment. Participants listen to brief passages and answer questions that target several dimensions of language comprehension (i.e., understanding of the main idea, memory of facts and details, recall of event sequences, and making inferences and predictions). Answers to questions are recorded and scored as correct or incorrect according to response norms provided by test developers. Seven raters independently double-scored a third of the audio-recordings. The overall interrater reliability was 93.6% for the 5 to 6 age form and 93.5% for the 7 to 8 age form. Discrepancies were resolved through discussion. Internal consistency was α = .81 for the 5 to 6 age form and α = .87 for the 7 to 8 age form. Students' scale scores were used in the analyses.

Reading comprehension. GMRT-IV (MacGinitie & MacGinitie, 1989) was used as a criterion measure in Grades 1 and 2. The GMRT consists of 10 passages. For Grade 1, there are seven narrative and three expository passages. For Grade 2, there are six narrative and four expository passages. Each passage is divided into short segments. Each segment is accompanied by a multiple-choice item consisting of three images. Students select the image that matches the content of the story. GMRT is group administered and limited to 35 min. Reliability for scores were KR-20 = .88 for Level 1 and KR-20 = .92 for Level 2. The Adaptive Reading measure (aReading; Christ et al., 2014; $\alpha > .91$ across Grades K–5) was used as a criterion measure in Kindergarten. aReading is a computer adaptive measure that evaluates concepts of print, phonological awareness, and decoding. Students' item response theory (IRT)-derived theta scores were used in the analyses.

Procedure

All students completed MIA and CELF-5-USP before and after an 8-week supplemental instructional program designed to train inference making in kindergarten (project ELCII) and Grades 1 and 2 (project TELCI). Students in kindergarten completed aReading, whereas students in Grades 1 and 2 completed the GMRT. All measures were administered in a quiet space by trained research staff.

Data Analysis and Results

Validity Evidence

First, we examined the item difficulty and discrimination of MIA items using classical test theory (CTT) to identify problematic items and select a final set of items that captures a wide range of abilities. Item difficulty is the proportion of students who respond to the item correctly. Item difficulty is higher when fewer students answer the item correctly. Item discrimination is the correlation between responses for an item and the total score, which indicates whether the item measures the same construct as the other items. A high-quality item should have a moderate to high correlation (>.20) with the total score (Everitt & Skrondal, 2010). As can be seen in Table S1 (see online supplemental materials), item difficulty values covered a wide range, though some items had low values, indicating that these items were generally difficult for most students. Particularly, items in *Cephalopods* were the most difficult. Items in *Granny's Glasses* and *Eagles* had similar difficulty. Items in *Apple Thieves* appeared to cover the widest range in difficulty.

Second, we used the Rasch model (Rasch, 1960) to calibrate MIA items and calculate item difficulty parameters and student abilities within the IRT framework using Winsteps (Linacre, 2019). Regarding item calibration results from the Rasch model, Table S2 (see online supplemental materials) shows the estimated item parameters and their standard errors. The Rasch model places both items and students onto the same logit scale. Figure S3 (see online supplemental materials) shows that the item difficulty values ranged from −2 to +2, and students' ability levels ranged from −3 to +2. Thus, students' ability levels aligned with the difficulty levels of the items. Figure S4 (see online supplemental materials) shows that MIA was highly informative (i.e., precise in terms of measurement) within (−2 to +2) range of student ability, and the amount of conditional standard error of measurement (cSEM) was quite low within this range. In addition, the item set is well discriminating within this range. Overall, these findings suggest that the difficulty level of MIA was suitable for this sample of students in K–2.

Third, we evaluated construct validity evidence. The results of the Rasch model indicated that a unidimensional vertical scale indeed had a good fit to the data and explained 40% of the total variance. We also used CFA to confirm the unidimensional structure. The CFA results indicated that the assessment had acceptable levels of model-data fit (comparative fit index [CFI] = .90, Tucker–Lewis index [TLI] = .90, root mean square error of approximation [RMSEA] = .023). These findings suggest that there is adequate evidence supporting the unidimensionality of MIA.

Finally, we examined criterion-related validity evidence. To do so, we computed bivariate correlations among MIA and measures of language comprehension and reading. Consistent with other nonreading inference-focused measures (e.g., LARRC Inference Task), we expected weak to moderate positive correlations. Indeed, correlations between MIA and CELF-5-USP ranged from $r = .322$ to .403 ($p < .01$) in K–2; MIA and GMRT from $r = .326$ to .339 ($p < .01$) in Grades 1 and 2; and MIA and aReading from $r = .216$ to .385 ($p < .05$) in K.

Reliability/Precision Evidence

We examined the reliability and precision of MIA scores based on a variety of indices. Particularly, the coefficient alpha value was .88, suggesting that the assessment had good internal consistency. Alternative reliability indices were also evaluated: Guttman L2 = .88, Feldt-Brennan = .88, and Feldt-Gilmer = .88. The Person separation index from the Rasch model was .85, which also suggests that MIA produces reliable scores.

Intended Use of Scores

We evaluated the adequacy of four parallel, equivalent forms with fewer items, to provide evidence for instructional sensitivity. To create the forms, we implemented an ATA procedure, which uses computer algorithms and mathematical optimization techniques to construct parallel test forms that satisfy a set of psychometric, content, and test administration constraints. For MIA, the parallel test forms were developed based on the following constraints: (a) each form will consist of 16 items, (b) each item can be used only once across the forms, (c) each form will include items from each module (*Apple Thieves* or *Granny's Glasses; Cephalopods* or *Eagles*), and (d) each form will maximize the test information (i.e., measurement precision) within the ability range of −2 to +2. ATA was implemented with the *xxIRT* package (Luo, 2019) in R (R Core Team, 2019). The ATA procedure was able to solve the optimization problem and yielded an optimal solution based on the specified constraints. See Table S3 (see online supplemental materials) for a summary of the four test forms. The forms yielded very similar levels of internal consistency ($\alpha_{Form1} = .87$, $\alpha_{Form2} = .87$, $\alpha_{Form3} = .90$, $\alpha_{Form4} = .90$).

Next, we used the scores from these forms to evaluate instructional sensitivity. Students in K–2 received inference instruction for 8 weeks, and scores on these forms were used to evaluate sensitivity to instruction from pretest to posttest. There was no control group in this proof-of-concept study. Rather, the focus was to evaluate whether MIA would be sensitive to monitor and evaluate progress of inference-focused instruction. *Effect sizes* varied from *medium* (Form 1: std. beta = 0.62, std. *SE* = 0.08, $\eta^2 = .15$; Form 3: std. beta = 0.63, std. *SE* = 0.08, $\eta^2 = .20$,) to *small* (Form 2: std. beta = 0.46, std. *SE* = 0.08, $\eta^2 = .13$; Form 4: std. beta = 0.38, std. *SE* = 0.08, $\eta^2 = .09$) showing promise for MIA to be used for *progress monitoring* during inference-focused instruction.

Discussion and Implications

The overall aim of the current investigation was to develop and validate the initial version of MIA. MIA is a web-based measure of inference processes in K–2. MIA leverages the affordances of different media to evaluate inference processes in a nonreading context using age-appropriate fiction and nonfiction videos coupled with questioning. We evaluated MIA's technical adequacy in a proof-of-concept study.

Taken together, the results support the interpretation that MIA shows promise as a valid and reliable measure of inferencing in a nonreading context for students in K–2. MIA has a strong theoretical basis, conceptualizing inference processes as unidimensional and independent of media factors (Kendeou et al., 2020). The validity argument for the current version of MIA shows that the underlying construct of inference processes assessed is *unidimensional*, has moderate correlations with measures of language and reading comprehension, and is with a difficulty level suitable for students in K–2. These findings are consistent with those obtained for other nonreading inference measures

(LARRC & Muijselaar, 2018). Reliability/precision evidence for MIA and its parallel forms was also adequate.

Despite the considerable strengths of MIA, this initial version has several limitations that need to be addressed with further development and refinement. First, the current four parallel forms use some of the same videos. Watching the same video twice may influence performance due to familiarity. Thus, we need to increase the pool of videos (and questions) to create unique, parallel forms suitable for progress-monitoring purposes. Second, although the current sample was ethnically and economically diverse, students from only one school district in the midwestern United States were included with a small number of students in Grade 2, thereby limiting the generalizability of the current findings. Future work should include large and diverse samples in K–2 across the United States; because English Learner status was not available in the current sample, it is also important to evaluate the appropriateness of MIA for English Learners in future studies. Finally, progress monitoring and instructional utility of MIA scores were only evaluated at two time points, pretest and posttest, in the context of an 8-week supplemental instruction on inference processes. In future studies, we need to evaluate the use of the parallel forms at multiple time points and in relation to a control group to establish technical adequacy for both progress monitoring and efficacy of instruction.

We contend that further development and refinement of MIA has the potential to produce a scalable, web-based assessment with fully automated test administrations and score reports that can help researchers and teachers evaluate inference processes independent of decoding in the early years.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The research reported herein was funded by grant numbers R324A160064 and R305A170242 from the U.S. Department of Education to the University of Minnesota. The opinions are those of the authors and do not represent the policies of the U.S. Department of Education.

ORCID iD

Reese Butterfuss https://orcid.org/0000-0001-9326-4176

Supplemental Material

Supplemental material for this article is available on the *Assessment for Effective Interventions* website with the online version of this article.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. AERA.

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.

Barth, A. E., Barnes, M., Francis, D., Vaughn, S., & York, M. (2015). Inferential processing among adequate and struggling adolescent comprehenders and relations to reading comprehension. *Reading and Writing*, *28*(5), 587–609.

Biancarosa, G., Kennedy, P. C., Carlson, S. E., Yoon, H., Seipel, B., Liu, B., & Davison, M. L. (2019). Constructing subscores that add validity: A case study of identifying students at risk. *Educational and Psychological Measurement*, *79*(1), 65–84.

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing*, *11*(5–6), 489–503.

Carlson, S. E., Seipel, B., & McMaster, K. (2014). Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learning and Individual Differences*, *32*, 40–53.

Christ, T. J., Aranas, Y. A., Kember, J. M., Kiss, A. J., McCarthy-Trentman, A., Monaghen, B. D., . . .White, M. J. (2014). *Formative assessment system for teachers technical manual: EarlyReading, CBMReading, aReading, aMath, and earlyMath*. Formative Assessment System for Teachers.

Everitt, B., & Skrondal, A. (2010). *The Cambridge dictionary of statistics*. Cambridge University Press.

Graesser, A. C. (2015). Deeper learning with advances in discourse science and technology. *Policy Insights from the Behavioral and Brain Sciences*, *2*, 42–50.

Graesser, A. C., & Franklin, S. P. (1990). QUEST: A cognitive model of question answering. *Discourse Processes*, *13*(3), 279–303.

Kendeou, P. (2015). A general inference skill. In E. J. O'Brien, A. E. Cook, & R. F. Lorch, Jr (Eds.), *Inferences during reading* (pp. 160–181). Cambridge University Press.

Kendeou, P., Bohn-Gettler, C., White, M. J., & Van Den Broek, P. (2008). Children's inference generation across different media. *Journal of Research in Reading*, *31*(3), 259–272.

Kendeou, P., McMaster, K. L., Butterfuss, R., Kim, J., Bresina, B., & Wagner, K. (2020). The inferential language comprehension (iLC) framework: Supporting children's comprehension of visual narratives. *Topics in Cognitive Science*, *12*, 256–273.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Language and Reading Research Consortium. (2015). The dimensionality of language ability in young children. *Child Development*, *86*(6), 1948–1965.

Language and Reading Research Consortium, & Muijselaar, M. M. (2018). The dimensionality of inference making: Are local and global inferences distinguishable? *Scientific Studies of Reading*, *22*(2), 117–136.

Linacre, J. M. (2019, May). *Reliability and separation of measures*. https://www.winsteps.com/winman/reliability.htm

Luo, X. (2019). *xxIRT: Item response theory and computer-based testing in R* (R package version 2.1.2). https://CRAN.R-project.org/package=xxIRT

MacGinitie, W. H., & MacGinitie, R. K. (1989). *Gates–MacGinitie reading tests*. Riverside.

Magliano, J. P., Kopp, K., McNerney, M. W., Radvansky, G. A., & Zacks, J. M. (2012). Aging and perceived event structure as a function of modality. *Aging, Neuropsychology, and Cognition*, *19*(1–2), 264–282. https://doi.org/10.1080/13825585.2011.633159

Magliano, J. P., Radvansky, G. A., & Copeland, D. E. (2007). Beyond language comprehension: Situation models as a form or autobiographical memory. In F. Schmalhofer & C. Perfetti (Eds.), *Higher level language processes in the brain: Inference and comprehension processes* (pp. 379–391). Lawrence Erlbaum.

McMaster, K. L., Van den Broek, P., Espin, C. A., White, M. J., Rapp, D. N., Kendeou, P., . . . Carlson, S. (2012). Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learning and Individual Differences*, *22*(1), 100–111.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, *38*, 1–30.

McNamara, D. S., & Magliano, J. P. (2009). Towards a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 297–384). Elsevier Science.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedagogiske Institut.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Semel, E., Wiig, E., & Secord, W. (2003). *Clinical evaluation of language fundamentals–4 (CELF-4)*. PsychCorp

Wiig, E. H., Semel, E. M., & Secord, W. (2003). *CELF 5: Clinical evaluation of language fundamentals*. Pearson/PsychCorp