

An analysis of standardized reading ability tests: What do questions actually measure?

Abstract. This study examined what types of passage and individual item attributes of reading ability tests affect item difficulty. Ninety-six questions from the Gates-MacGinitie Reading Test for 7/9th graders were analyzed in terms of individual item characteristics and passage features that were considered to affect the item's difficulty. Whereas two passage features, word frequency and sentence length, influenced the item difficulty, none of the individual question characteristics were found to affect the item difficulty. Thus, the GMRT may be assessing how difficult a passage they can comprehend, not how difficult a question they can answer about a passage. These findings may be the result of the passages in the GMRT having an atypical construction; shorter passages tended to be composed of longer sentences.

Introduction

Research on reading comprehension indicates that successful comprehension of text requires the effective execution of a number of sub-processes such as word decoding, lexical access, syntactic processing, and constructing a coherent overall meaning of a text by connecting multiple sentence meanings (Vellutino, 2003). However, in most practical contexts, a student's reading comprehension ability is evaluated in an undifferentiated fashion by standardized tests such as the Gates-MacGinitie reading ability test (Magliano & Millis, 2001). In a typical reading comprehension ability test, students read a short passage and answer multiple-choice questions based on the passage content. This test format is considered a measure of reading comprehension ability because selecting the correct answer involves reading activities such as understanding the meanings of the passage and the question, and searching for a potential answer in the passage. Indeed, performance on these tests highly correlates with other types of standardized measures of academic ability (e.g., Comprehensive Test of Basic Skills, Preliminary Scholastic Aptitude Test, reading course grades, language course grades; see technical report of Gates-MacGinitie Reading tests, 1989).

Although the benefit of standardized tests is plainly evident, we, along with other researchers (e.g., Embretson, 1994), argue that it is time to move away from an a-theoretical approach of measuring reading ability such as those previously used. Advances in reading comprehension research have made it possible to analyze item characteristics of reading comprehension test questions in terms of their necessary cognitive processes. A better understanding of the relationship between question item characteristics and the cognitive processes relevant for answering them would facilitate development of more theoretically motivated reading comprehension assessment tools. Such assessment tools would be able to identify specific deficits or problems in readers' comprehension abilities (e.g., vocabulary, syntax). In sum, reading comprehension ability tests should not only identify who has a reading problem, but also diagnose what type of problem they have (Cornoldi, De Beni, & Pazzaglia, 1996).

Analyzing individual items of reading ability test

Embretson and Wetzel (1987) examined the item difficulty of several different reading comprehension tests using a number of different coding systems based on the cognitive processing model of reading comprehension tests. According to this model, answering reading comprehension questions involves two stages: 1) encoding a passage by forming a coherent representation of the text, and 2) making a decision concerning which answer option to choose. Encoding the passage is the construction of a coherent representation of the overall meaning of the passage by integrating textbase information and prior knowledge (Kintsch, 1988, 1998). The decision stage involves three components: 1) encoding of the question stem and answer options, 2) mapping and comparing the questions/answer options to the passage, and 3) evaluating the truth status of the answer options. According to this model, item difficulty is influenced by the difficulty of the processing required in each of the components. Specifically, a question item is expected to be more difficult when: 1) the passage contains more information, 2) the question requests more information, 3) overlap between the question/answer options and the passage content is small, and 4) answer options cannot be explicitly confirmed by the text content.

Based on these hypotheses, Embretson and colleagues (Embretson & Wetzel, 1987; Gorin, Embretson, & Sheehan, 2002) analyzed question items contained in several reading comprehension tests (e.g., Army Services Vocational Aptitude Battery & GRE) and showed that features of the passage and question items were related to item difficulty. More specifically, they found that the propositional density of the passage and the extensiveness of reasoning required to map the question/answer on the passage were two main factors that influenced item difficulty. A similar analysis was performed by Sheehan and Ginther (2001) for reading comprehension questions from a test of English as a Foreign Language. They found that item difficulty was related to the activation value of the information in the passage (e.g., frequency, the location of the target information). Overall, studies have shown that item difficulty of reading comprehension ability tests is related to text and/or individual item features in theoretically meaningful ways. Following this approach, we

explored whether the item difficulties of comprehension test items of the Gates-MacGinitie Reading Test (grade level 7/9; forms K and L) were related to some specific, theoretically important, item characteristics.

Present study

In this study we followed Embretson and colleague's (e.g., Embretson & Wetzel, 1987) approach, but also included some other theoretically motivated measures. We analyzed the passage difficulty in terms of several text features that are known to affect reading comprehension, including: word frequency, sentence length (number of words per sentence), argument-overlap between the adjacent sentences (Britton & Gulgoz, 1991), and the overall propositional density of the text. When a passage contains unfamiliar words, readers often experience difficulty understanding the text, resulting in an increased difficulty of the questions associated with the passage. Sentence length is also considered to affect processing demand; processing a longer sentence places larger demands on working memory (Just & Carpenter, 1992), potentially rendering the passage more difficult. Finally, less argument overlap between adjacent sentences places demands on the reader because the reader needs to infer the relations between the sentences to construct a global representation of the text (Britton & Gulgoz, 1991; Kintsch, 1998).

Item characteristics of individual questions are assumed to influence the decision stage in the answering process. To assess these characteristics, we used Embretson's coding scheme, which included: 1) reasoning required to identify the correct answer (Anderson, 1982), 2) confirmability of correct answer, and 3) the number of distractors that can be explicitly falsified. In addition, we also included: 1) the degree of inferential processing of the text required to answer a question, and 2) Mosenthal's (1996) coding scheme of abstractness of the requested information, specifically, the level of abstractness in the information requested by the question and target answer. Following Gorin et al. (2002), we expect that more difficult items tend to involve questions requiring more reasoning, more abstract information, or more inferences. We also expect that items tend to be more difficult when the target cannot be directly confirmed and/or distractors cannot be falsified based on the passage content.

Method

Materials

Ninety-six multiple choice questions from the comprehension portion of the Gates-MacGinitie reading test (GMRT) grade level 7/9, forms K and L, were the target of this analysis. The technical report for the GMRT third edition (Gates-MacGinitie Reading Tests Technical Report, 1989) contains item difficulty parameters in terms of the proportion of readers who answered the item correctly. This item difficulty score differs from item difficulty parameters based on Item Response Theory which was used in the Gorin et al. (2002) study. The technical report listed item difficulty for 7th, 8th, and 9th grade children separately for fall and spring semesters. We used item difficulty corresponding to spring 9th grade children for the current analysis. The first two practice questions for each form were excluded from the analysis. Two coders were trained on every coding scheme and rated 15% of total question items from 3 different texts within the GMRT. The two raters agreed at least 80% of the time across all the dimensions of coding. The remainder of the question items and passages were rated by both coders together.

Coding

Ninety-six individual questions and their answer options were coded in terms of their relationship with the target passage in five different ways. The first coding was on abstractness of the information requested by a question using Mosenthal's (1996) scheme. This score addresses the level of abstractness or concreteness of the information requested by the question. Abstractness of the item was classified using a five-point scale. The first classification of questions, *most concrete*, asks for the 'identification of persons, animals, or things.' The *highly concrete* class of questions ask for the 'identification of amounts, times, or attributes.' *Intermediate* questions ask for the 'identification of manner, goal, purpose, alternative, attempt, or condition.' In the *highly abstract* class, the questions

ask for the ‘identification of cause, effect, reason, or result.’ In the *most abstract* class, the questions ask for the ‘identification of equivalence, difference, or theme.’

The second coding scheme assessed the degrees of inferential processes necessary to answer questions by examining the degree of overlap between the item (e.g., question and target answer) and the passage. *Text-based* items involved questions and answers that can be found in a single sentence, and the item and passage content were worded the same. *Reworded* items were similar to *text-based* items but items were worded with synonyms. *Inference* items were those in which the correct answer could not be directly found in the passage. Finally, *integration* items were those that required integration of multiple sentence meanings from a passage to answer.

The third coding system involved application of the Anderson’s (1982) rating system used by Embretson and Wetzel (1987). This score addressed the levels of transformation required to match a question item to the passage. There are four levels in this scheme. The lowest level, *verbatim*, uses the exact wording used in the passage. The next level is *transformed verbatim question*, in which the same words are used in the question as in the passage, but the sentences are rearranged. In *paraphrased questions*, the question has the same meaning as a sentence in the passage, but different words are used. In *transformed paraphrase questions*, neither the wording nor the phrase order in the question were the same as in the passage.

In addition to the above three schemes that analyze question stem and correct answer together in relation to the passage, we also coded the quality of answer options in two different ways following Embretson and Wetzel (1987). The first score addressed whether the correct answer could be directly confirmed by the passage content. A target answer is *confirmed* if the passage provides explicit textual evidence that the target is the correct answer. However, a target answer is *not confirmed* if the passage has no mention of the target answer. A second score addressed how many of the distractors could be explicitly falsified by the content of the passage. A distractor is *falsifiable* if the passage provides explicit textual evidence that the distractor is incorrect.

Passage features were analyzed using Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004), a computational tool that assesses 238 cohesion and readability measures that are considered to influence comprehension. For the current study, we focused on five measures that are either theoretically or empirically known to influence comprehension difficulty, including: Flesch reading ease, number of propositions, number of words per sentence, word frequency, and argument overlap between adjacent sentences. Number of propositions was calculated using the Charniak parser (Charniak, 2000) to count the number of subject level parsings.

Results

Before presenting the results of the detailed analysis on the influence of the individual items and passages on item difficulty, we present the overall, average passage features of the GMRT in Table 1. As indicated, the GMRT is comprised of passages that vary in difficulty from 5th to 12th grade level texts. Also, the data indicate that overall average sentence length is excessive, with the longest sentence having as many as 44 words. These forms of the GMRT had more questions based on narrative ($N = 51$) passages as opposed to expository ($N = 40$) passages. The remaining items ($N = 5$) were based on poetry.

Table 1 Overall passage features used in the GMRT form K and L

| | Mean | SD | Min/Max |
|-------------------|------|------|------------|
| Flesch Kincaid | 7.9 | 2.40 | 4.79/12.00 |
| Word per sentence | 19.2 | 8.40 | 8.40/44.50 |
| Number of words | 99.0 | 22.0 | 58/130 |

The results of the main analyses are described in two sections: 1) the effect of the nature of individual questions (question stem and answer options) on item difficulty, and 2) the effect of

passage characteristics on item difficulty. Item difficulty is measured in terms of performance (proportion correct) reported in the GMRT manual for the spring 9th grade children. The analysis is split into two sections because passages and question items are affected by different processes. Being able to answer questions about a passage requires more than simply comprehending the passage. That is, information from the passage can be requested in a question item in a number of different ways.

The nature of individual items (question stem and answer options)

First, we analyzed the item difficulty in terms of the nature of individual question items. As described in the method section, we used three types of coding to identify the nature of requested information by question items: 1) Mosenthal’s (1996) abstractness of the requested information, 2) degree of inference needed about the passage to answer the question item (e.g., text-based, paraphrased, inference, integration of multiple sentences), and 3) how close the wording and the structure of an item (question and correct answer) is to the content of the target passage (adapted from Anderson, 1982). Table 2 indicates the item difficulty parameter as a function of Mosenthal’s classification of information abstractness. A one-way ANOVA indicated that there were no significant differences in item difficulty among the five types of requests for information, $F(4, 91) < .1$.

Table 2: Item difficulty as the function of information abstractness

| | Most concrete | Highly concrete | Intermediate | Highly abstract | Most abstract |
|------|---------------|-----------------|--------------|-----------------|---------------|
| N | 9 | 34 | 18 | 23 | 12 |
| Mean | .65 | .66 | .72 | .66 | .65 |
| SD | .12 | .16 | .15 | .16 | .17 |

Table 3 shows item difficulty as the function of the degree of inference necessary to answer questions about a passage, in other words how much the question is changed from the passage. A one-way ANOVA indicated that there were no differences in item difficulty among the four types of questions, $F(3, 95) < .1$.

Table 3: Item difficulty as the function of degrees of inference necessary to answer the questions

| | Text-based | Re-wording | Inference | Bridging |
|------|------------|------------|-----------|----------|
| N | 23 | 26 | 30 | 17 |
| Mean | .67 | .66 | .67 | .67 |
| SD | .15 | .14 | .17 | .16 |

Finally, we examined whether item difficulty was affected by the match between the question stem and the passage content using Anderson’s classification scheme. Table 4 indicates item difficulty as a function of Anderson’s classification of question transformation. A one-way ANOVA was performed (excluding the verbatim items due to small representation in the sample; $n = 1$). The analysis indicated no difference in item difficulty among the three kinds of items, $F(2, 95) = 1.292$, $MSE = .024$, $p > .2$.

Table 4: Item difficulty as the function matching between the stem and passage

| | Verbatim | Transform Verbatim | Paraphrase | Transform Paraphrase |
|------|----------|--------------------|------------|----------------------|
| N | 1 | 13 | 41 | 41 |
| Mean | .59 | .68 | .64 | .71 |
| SD | - | .11 | .17 | .15 |

Overall item analyses with three different types of coding based on the processing required for answering questions indicated that the nature of the question item is not the primary factor moderating item difficulty in GMRT (7/9th grade) forms K and L.

Following the classification system of Embretson and Wetzel (1987), we analyzed whether confirmability of the target answer and/or falsifiability of distractors based on explicit passage content affected item difficulty. An examination of item difficulty as a function of confirmability of the target answer indicated that there was no difference in item difficulty between items with confirmable target answers ($M = .67$, $SD = .15$) and items with target answers that are not confirmable ($M = .67$, $SD = .15$), ns. To examine whether the falsifiability of distractors affected item difficulty, we correlated the number of falsifiable distractors (i.e., 0 to 3 answer options) and item difficulty, and found that there was almost no correlation between them ($r = .050$, ns). These results indicated that confirmability and falsifiability of answer options for multiple choice questions do not influence item difficulty in the 7/9th grade GMRT.

The nature of passage characteristics

We correlated item difficulty with several key passage features derived from Coh-Metrix. The correlations are presented in Table 5. The correlations reveal several interesting findings. First, word frequency minimum, a measure representing average lowest frequency word per sentence within each passage, has a significant, positive correlation with item difficulty, indicating that students tend to perform poorly when a question was based on a passage with sentences having more low frequency words overall. When the different types of texts were examined separately, the minimum word frequency was significantly correlated with item difficulty for expository passages ($r = .326$, $p = .43$), but not for narrative texts ($r = .15$, ns.). The average frequency of words for an entire passage was not found to be significant. This finding indicates that the comprehension of a sentence is more closely related to how unfamiliar the least frequent words in a sentence are, as opposed to the overall word frequency level. For example, a sentence containing four frequent words and a single very, unfamiliar word is often more difficult to understand than a sentence containing five moderately frequent words.

Second, a significant, negative correlation between item difficulty and the average number of words per sentence for an entire passage suggested that items based on a passage with long sentences tended to be more difficult. A significant, positive correlation between item difficulty and Flesch reading ease is consistent with these findings because reading ease is calculated based on sentence length and word frequency (i.e., based on word length).

Adjacent argument overlap, an index of local text cohesion, was negatively correlated with item difficulty, suggesting that passages with greater local cohesion tended to be associated with more difficult questions. This finding is contrary to the predicted direction of this correlation, and suggests that a passage with greater co-referential scaffolding between adjacent sentences tends to be more difficult, or at least the questions that target passages with more overlap are more difficult. Finally, contrary to Embretson and Wetzel (1987), number of propositions was found to be positively correlated with items difficulty, suggesting that questions based on a propositionally dense passage tended to be easier in GMRT for 7/9th grade level.

Table 5: Correlation between key passage features and item difficulty parameter

| | Item difficulty | Number of propositions | Words per sentence | Reading ease | Word frequency |
|------------------------------------|--------------------|------------------------|--------------------|--------------------|--------------------|
| Number of propositions | .315** p < .01 | | | | |
| Number of words per sentence | -.265** p < .01 | -.397** p < .01 | | | |
| Flesch reading ease | .269** p < .01 | .287** p < .01 | -.791** p < .01 | | |
| Word frequency (min each sentence) | .325** p < .01 | .258* p < .05 | -.754** p < .01 | .738** p < .01 | |
| Adjacent argument overlap | -.216* p < .05 | -.116 | .598** p < .01 | -.628** p < .01 | -.480** p < .01 |

Discussion

This study examined whether question and passage attributes of the Gates-MacGinitie Reading Test for 7/9th graders affected item difficulty in the 96 questions from the test. We analyzed the questions in terms of individual item characteristics and passage features. We found that none of the individual item features (e.g., abstractness of the requested information, matching between question and passage) were found to be related with the item difficulty in a systematic way. There are two possible explanations for this result. First, the coding schemes may be incorrect. Second, the factors assessed by the coding schemes are not significant predictors for these GMRT item difficulty ratings; and other factors determine the differences in item difficulty. The first possibility is unlikely because the items were coded consistently following the instructions provided by the authors who created the coding schemes (e.g., Mosenthal, 1996). Furthermore, at least two of the coding schemes correlated with each other (Mosenthal & Anderson, $r = .347$, $p < .05$), showing that some questions are consistently more cognitively demanding than others. Thus, the null effect is more likely to be attributable to the presence of other factor(s) that override the influence of individual question features. One possibility is that these coding schemes are designed to differentiate items based on the involvement of relatively higher level processing (e.g., inference, transformation, and construction of overall understanding). It may be the case that student performance is primarily differentiated by their lower level processing proficiency (e.g., vocabulary, syntax), making lower level processing a more prevalent moderator of GMRT item difficulty.

In line with the above interpretation, item difficulty was correlated with minimum word frequency and sentence length. A negative correlation was found between item difficulty and minimum word frequency, suggesting that questions targeting passages with more common words are easier to answer. However, this correlation was only found in expository passages suggesting the involvement of different factors in narrative and expository passage comprehension. The negative correlation between sentence length and item difficulty, which is in the expected direction, suggests that it is more difficult to answer questions based on passages composed of longer sentences. These findings are generally in line with reading comprehension literature (Vellunito, 2003).

Interestingly, two passage features were correlated with item difficulty in theoretically unexpected directions. Number of propositions was expected to negatively correlate with item difficulty because searching for an answer in a passage with a larger number of propositions is more demanding. Argument overlap of neighboring sentences was expected to positively correlate with item difficulty because larger argument overlap facilitates the construction of a global meaning of a text by providing extra cues to connect sentences. Contrary to these expectations, the results showed positive correlation between item difficulty and a number of propositions and negative correlation between argument overlap and item difficulty. These two unexpected findings may be due to peculiar features of the passages used in the GMRT; the data shows shorter passages tended to be composed of longer sentences (significant negative correlation between overall propositional density and number of words per sentence). In the most extreme example, a passage of 90 words only had two sentences. Longer passages were composed of more sentences which were shorter in length, making longer passages relatively easier to comprehend than shorter passages composed of longer sentences. Also, the data indicated that longer sentences are more likely to have a larger argument overlap (significant positive correlation between argument overlap and the number of words per sentence). This latter phenomenon is also understandable because having longer sentences increases the probability of having common words between them. Now, taking these factors into account, the positive correlation between number of propositions and item difficulty is a possible by-product of the presence of numerous propositions in short sentence passages. Likewise, the negative correlation between argument overlap and item difficulty could also be due to sentence length.

Overall, item difficulty of the questions in the GMRT for 7/9th grade appears to be primarily related to relatively lower level processes, namely vocabulary and the ability to process long and complex sentences. The primary dependence of item difficulty on lower level processes is not a good or bad thing in itself. Yet, the current finding poses a question about the use of GMRT in assessing students' reading ability. That is, the finding indicates that GMRT differentiates high and low ability

students based on how difficult a passage (with low frequency word and long sentences) students can read, not how deeply (e.g., answering more demanding questions based on a passage) students can process a given passage. Thus, whether the GMRT is appropriate may depend on the type of reading ability one would like to assess.

References

- Anderson, R. C. (1982). How to construct achievement tests to assess comprehension. *Review of Educational Research*, 42, 145-170.
- Britton, B. K., & Gulgoz, S. (1991) Using Kintsch's computational model to improve instructional text: effect of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, 83, 329-345.
- Charniak, E. (2000) A maximum-entropy-inspired parser. In Proceedings of NAACL-2000. pp. 132–139.
- Cornoldi, C., De Beni, R., & Pazzaglia, F. (1996). Profiles of reading comprehension difficulties: An analysis of single cases. In C. Cornoldi & J. Oakhill (Eds.), *Reading Comprehension Difficulties: Processes and Intervention*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Embretson, S. E. (1994) Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive Assessment: A multidisciplinary perspective* (pp 107-135). New York: Plenum Press.
- Embretson, S. E. & Wetzel, C. D. (1987). Component latent models for paragraph comprehension, *Applied Psychological measurement*, 11, 175-193.
- Gates-MacGinitie Reading Tests (1989). Technical report for Gates-MacGinitie Reading Tests Forms K and L. Chicago, IL: Riverside Publishing
- Gorin, J. S., Embretson, S. E., & Sheehan, K. (2002). *Cognitive and psychometric modeling of text-based reading comprehension GRE-V items*. Paper presented at the 2002 Annual Meeting of the National Council on Measurement in Education: New Orleans, LA.
- Graesser, A.C., McNamara, D. S. Louwerse, M., & Cai, Z. (2004). Coh-Matrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments and Computers*, 36, 193-202.
- Just, M. A., & Carpenter, P. A. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99, 122-149.
- Kintsch, W. (1998). *Comprehension: A Paradigm for cognition*. Cambridge, MA: Cambridge University Press.
- Magliano, J. P. & Millis, K. (2003). Assessing reading skill with a think aloud procedure and latent semantic analysis. *Cognition & Instruction*, 21, 251-283.
- Mosenthal, P. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology*, 88, 314-332.
- Sheehan, K. M. & Ginther, A. (2001). *What do passage-based multiple choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section*. Paper presented at the 2000 Annual Meeting of the National Council of Measurement in Education.
- Vellutino, F. R. (2003). Individual differences as sources of variability in reading comprehension in elementary school children. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 51-81) New York: Guilford Press.

Acknowledgements

The research was supported by the Institute for Education Sciences (IES R3056020018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES.