Is the Treatment Weak or the Test Insensitive: Interrogating Item Difficulties to Elucidate the Nature of Reading Intervention Effects

David J. Francis, Paulina A. Kulesz, Shiva Khalaf, Martin Walczak

Department of Psychology

and

Texas Institute for Measurement, Evaluation, and Statistics

University of Houston

Sharon R. Vaughn

Department of Special Education

and

Meadows Center for the Prevention of Educational Risk

The University of Texas at Austin

Acknowledgements: The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305F100013 to The University of Texas at Austin as part of the Reading for Understanding Research Initiative, and through Grant R305A170251 to the University of Houston under the Education Research program of the National Center for Education Research (NCSER), and the Eunice Kennedy Shriver National Institute of Child Health and Human Development P50 HD052117, Texas Center for Learning Disabilities. The opinions expressed are those of the authors and do not represent views of the Institute, NCSER, the U.S. Department of Education, NICHD, or the National Institutes of Health.

Declarations of Interest: None

Please address all correspondence to:

David J. Francis, Ph.D.
Director, Texas Institute for Measurement, Evaluation, and Statistics
HEALTH-1, Room 372
4349 Martin Luther King Boulevard
Houston, TX. 77204-6022

Accepted for publication in Learning and Individual Differences, Volume 97, 2022. Pages not yet available.

Abstract

Intervention research in education is sometimes criticized for the use of experimenter developed assessments, especially when these are over aligned with treatment. At the same time, intervention researchers sometimes prefer locally developed assessments because they appear to be more sensitive to treatment effects even when the test is not subject to the criticism of over alignment. This paper examines the question of test sensitivity to treatment effects for experimenter developed and standardized tests for the specific case of reading in grade 8. We examine similarities and differences between a specific experimenter developed test and widely used standardized reading assessment. Analyses show these particular tests to be quite comparable. The paper concludes with an examination of test sensitivity by simulating treatment effects of different magnitudes. These analyses highlight some potential limitations of the standardized test for detecting small to moderate effects depending on the ability range of the students participating in intervention. The implications for intervention research and identification of students under response to intervention are discussed.

Is the Treatment Weak or the Test Insensitive: Interrogating Item Difficulties to Elucidate the Nature of Reading Intervention Effects

Intervention Effectiveness on Researcher-Developed and Standardized Assessments

The What Works Clearinghouse and Best Evidence Encyclopedia both aim to provide researchers and educators with unbiased syntheses of intervention/treatment effects (Slavin, 2008; Slavin & Madden, 2011). Yet, the magnitude of intervention effects (ES) widely varies depending on an assessment being used (researcher-developed vs standardized) to evaluate intervention outcomes (Cheung & Slavin, 2015). Numerous reading intervention studies involving elementary and middle school monolingual or bilingual students have found more positive ES (suggesting intervention effectiveness) on researcher-developed relative to standardized assessments. To exemplify: (a) the 2013 What Works Clearinghouse review on interventions for beginning reading reported higher ES for researcher-developed measures (mean ES = +2.49) than for standardized measures (mean ES = +0.93); (b) Vaughn and colleagues (2013) review of middle school interventions for reading comprehension found more positive ES on researcher-developed reading assessment (ES = +0.29) relative to a standardized assessment (ES = +0.19); (c) Kim and colleagues (2006) found higher ES on two researcher-developed measures (ES= +0.95 and +0.77) compared to a standardized measure (ES= +0.50); and (d) Boyle (1996) reported higher ES for researcher-developed measures (ES = +0.86) relative to standardized measures (ES = +0.33). Finally, Vaughn and colleagues (2017) found statistically significant differences between treatment and control groups on a researcher-developed measure of reading comprehension (ES = +0.20), though they did not find similar effects when using a standardized assessment of comprehension. While the examples are not exhaustive, they demonstrate a pattern of relatively greater ES in favor of researcher-developed measures that has

been observed elsewhere (Gonzalez et al., 2010; Mitchell & Fox, 2001; Powell, Diamond, Burchnal, & Koehler, 2010).

Alignment of Researcher-Developed and Standardized Assessments with an Intervention

Given this pattern of findings, it is natural to ask why interventions are generally more effective when assessed with researcher-developed assessments, or phrased differently, why standardized assessments are less sensitive to intervention effects. At least four explanations have been offered by researchers, three of which can be conceptualized in terms of alignment: (1) alignment of the intervention with the assessment, (2) alignment of the conceptual framework with assessments, and (3) a psychometric perspective on alignment/misalignment across assessments. The fourth explanation is related to measurable differences in the demands that assessments place on readers (Best et al., 2008; Cutting & Scarborough, 2006; Francis, Fletcher, Catts, & Tomblin, 2005; Garcia & Cain, 2014; Keenan, Betjemann, & Olson, 2008; Keenan & Meenan, 2014; Kendeou, Papadopoulos, & Spanoudis, 2012; Nation & Snowling, 1997).

Alignment of the intervention with assessment. In the 2012 commentary "Bad Science I: Bad Measures", Slavin argued that the problem of differential effect sizes stems from overly narrow alignment of researcher-developed measures with an intervention. Standardized assessments measure the construct of reading in a general way that neither advantages nor disadvantages groups, or individuals given prior background knowledge or experiences. These measures are not specifically designed to assess intervention effects (see Kulesz et al., 2016). At the same time, standardized reading measures assess the construct targeted by reading comprehension interventions, and as such, potentially measure the same construct as researcher-developed measures of reading. In contrast to standardized measures, researcher-developed measures are often designed to measure specific constructs, strategies, or topics that are the focus

of intervention, which leads to close alignment between the assessment and intervention and can give an unfair advantage to intervention participants, if controls lack exposure to the tested content. When topics covered in an intervention are assessed directly, the ES misrepresents the effects of intervention.

Alignment of conceptual framework with assessment. A second source of measurement-based heterogeneity in ES is differences in measures' conceptual frameworks (Best et al., 2008; Cutting & Scarborough, 2006; Francis et al., 2005; Garcia & Cain, 2014; Keenan, Betjemann, & Olson, 2008; Keenan & Meenan, 2014; Kendeou, Papadopoulos, & Spanoudis, 2012; Nation & Snowling, 1997). Specifically, when it comes to the conceptual frameworks for an assessment and the component skills that are assessed, Nation and Snowling (1997) found that when children are given a cloze format reading comprehension assessment along with measures of listening comprehension and decoding skills, students' decoding skills, not listening comprehension, explain a significant portion of variance in reading comprehension. Nation and Snowling's findings suggest that the cloze format reading comprehension assessment primarily measures students decoding skills rather than their language comprehension skills. Francis et al. (2005) found that decoding skills were more strongly associated with reading comprehension on cloze-based assessments than assessments using passages and multiple-choice question to assess comprehension. Under such circumstances, an intervention targeting decoding skills would be expected to yield different effects on a cloze-based and multiple-choice measures of reading comprehension.

A psychometric perspective on alignment across assessments. Many design features affect the sensitivity of a specific measure of reading comprehension to the effects of intervention. Standardized measures of comprehension are designed to measure ability over a

broad range (4 to 6 standard deviations of ability over one or more grade levels) using a relatively small set of items (35-45). Comparatively, interventions are often geared toward students reading in a narrow range of this ability continuum. A researcher-developed assessment designed to measure treatment efficacy might reasonably be expected to target a more narrow range of ability, providing more items within the range of ability expected among intervention participants. In other words, when items come from different assessments, treatment effect comparisons across assessment are complicated by differences in scale sensitivity, or the number of items available to assess ability within a given range of ability. With proper psychometric work using both assessments and a common sample, it is possible to place all items on a common scale, but such work is not routinely done in intervention research. Placing items on a common scale allows meaningful comparisons across different measures.

To exemplify, consider a student with a Lexile ability level of 800L reading two different, 40-item tests. Test W measures abilities ranging from 400L to 1200L on the Lexile scale and Test N measures abilities ranging from 650L to 950L. For ease of comparison, imagine that each test's items are spread uniformly across the ability range the test is designed to measure and that the test items are not otherwise different. Thus, the items differ only in their difficulty and the tests differ only in the range of abilities measured. If, following intervention, the student's ability increased from 800L to 850L, we would expect both tests to reflect this improvement, but test W is less sensitive to detecting the change because test W has half as many items within any score interval between 650L and 950L as test N. Of course, test N is insensitive to change below 650L or above 950L. Clearly, tests' sensitivity to changes in ability within the range of the participants' starting ability levels affects those tests' suitability for estimating intervention effects .

Measurable Differences in the Demands that Assessments Place on the Reader

Differences in the demands that assessments place on the reader are partially a function of the text being read (frequently a passage) and its textual features. Such effects have been widely examined within the text discourse framework (TDF) of reading. Interventions in the TDF have focused on engaging readers in recognizing and understanding the discourse features of the text and on developing and improving argumentation skills in the reader (Meyer & Ray, 2011; Pyle et al., 2017). Yet, the TDF has not been used to examine differences in intervention effectiveness on researcher-developed and standardized assessments, especially when the interventions are rooted within the component skills framework. Given the potential impact of text features on reading comprehension, it is plausible that differences in intervention effects on researcher-developed and standardized assessments arise at least partially from differences in text features. Keenan et al. (2008) compared four reading comprehension tests that used sentence-length passages (the Peabody Individual Achievement Test and the Woodcock–Johnson Passage Comprehension-3) and longer passages (the Gray Oral Reading Test-3 and the Qualitative Reading Inventory-3) and found that individual differences in reading comprehension tests were largely accounted for by word decoding skill as a function of passage length. Namely, longer passages required higher-levels of language skills when constructing mental models of situations that change across the sentences in a passage. Apart from passage length, other text characteristics demarcating text difficulty, such as word frequency, cohesion, and genre, have been found to affect reading comprehension. Texts with more frequent words and shorter sentences are generally easier to process because they require less cognitive effort and language ability (Abedi, 2006; Just, Carpenter, & Woolley, 1982; Perfetti, 1985; Turner, Valentine, & Ellis, 1998). Similarly, more cohesive texts place fewer demands on the reader because (by-and-

large) more cohesive passages do not require inference making and place weaker demands on readers' background knowledge when readers construct a mental story (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996; McNamara, Louwerse, McCarthy, & Graesser, 2010). Lastly, narrative texts that use more frequent words, have more connectives, and depend less on background knowledge are easier to comprehend (Linderholm & van den Broek, 2002; Vellutino, 2003).

Objectives

When treatment effects vary across assessments of the same construct, the assessment serves to moderate treatment effects. The question to be answered is why? If two tests measure the same construct and the intervention influences the construct, then treatment effects should be present on both instruments commensurate with their relationships to the construct and the ability ranges targeted by the intervention and the assessments. When the construct in question is reading comprehension, it is well known through the TDF (Francis, Kulesz, & Benoit, 2018; Kulesz, et al., 2016) those features of the text that affect the reader's comprehension.

Investigating intervention effects through the influence of text features on item difficulties across different assessments speaks directly to the treatment's mechanism of action. In depth interrogation of the effects of test construction measured via text features on students' performance across different assessments provides useful information for understanding the potential moderation of treatment effects by assessments. Does the treatment affect the construct, or sources of test-specific variation unique to a particular test or type of test, or does the intervention moderate the effects of specific text features on comprehension and these features are unequally represented on different tests? These scenarios suggest that tests can display differential sensitivity to treatment effects that do not stem from the over alignment. While over

alignment of tests to treatments is a problem for intervention research, it is not the only basis for differential test sensitivity. To date, no empirical studies have been conducted that seek to *empirically explain* the differential sensitivity of researcher-developed and standardized assessments to reading intervention treatment effects, such as through differential influences of text features on item difficulty. Although studies have speculated on the basis for differential test sensitivity, relying mostly on face validity evidence, deep empirical investigations have been conspicuously absent.

The current study investigates the role of text features on item difficulty in the context of testing the efficacy of a specific reading comprehension intervention in an effort to directly interrogate each test's sensitivity to change following intervention. Placing reading comprehension tests under the microscope can help to develop a clearer picture of precisely how tests differ from one another and why tests differ in their sensitivity to intervention effects, in other words, how test construction measured via text features may affect the test's sensitivity to treatment effects.

Specific answer(s) to this question may differ from one intervention study to the next and may well vary across standardized reading assessments that use different assessment frameworks. Consequently, our study does not seek to answer these questions in a general sense. Rather, we seek to develop data analytic steps to serve this purpose and to apply the methodology to data from a randomized experiment of an intervention designed to impact the reading comprehension of 8th Grade Social Studies students. Thus, the current paper examines the role of text features on item difficulty for different tests of reading comprehension within the context of evaluating the effectiveness of a specific reading comprehension intervention in a particular context, while also laying important groundwork for future development and

application of this methodology to other assessments and interventions.

Applied in an intervention context, the approach should be capable of distinguishing treatment effects from testing effects, while also signaling the extent to which treatment effects generalize across the universe of potential text passages and test items. The paper will demonstrate the importance of aligning researcher-developed and standardized tests before examining effects of interest across different assessments. To the extent that text feature effects are present, but do not interact with test type (researcher-developed and standardized), then we would expect that intervention effects would generalize across tests that are comparable on those text features, without regard to test type. Under such circumstances, the failure of treatment effects to generalize may indicate that tests are not equated for item difficulty and/or text features. To the extent that the features of text interact with test type to affect item difficulty, then intervention effects would not generalize across tests even when tests are equated on these text features. Under the latter circumstances, we would expect intervention effects to vary across researcher-developed and standardized tests both because text features differ across the two test types, but also because text features affect comprehension differently across tests. Through the equating of test items across researcher-developed and standardized tests and the modeling of text features on item difficulty it becomes possible to disentangle weak intervention effects from lack of sensitivity in the test.

Method

Participants

Table 1 presents demographic information for the total sample included in the current study (N = 1,957), and broken down by intervention status, Treatment (n = 818) vs. Control (n = 1,139). Eighth grade students were recruited from seven diverse middle schools located in five

large school districts in the Southeast and Southwest USA. At the time of the study (2011-2012, n = 644 and 2012-2013, n = 1,313) students were enrolled in US History classes (there was no student overlap across study years). Of the 1,957 students (male=942) who consented to participate in the study, 31 % qualified for free or reduced lunch. Students' average age was 13.16 in both conditions. Pretest mean scores (with standard deviations) for treatment and control conditions on the outcome measures (the Gates–MacGinitie reading comprehension subtest form T grade level 7/9, and Assessment of Social Studies Knowledge) are reported in Table 2. No significant differences between conditions on either of the measures were found at pretest.

[INSERT TABLES 1 AND 2 HERE]

Description of Intervention

Vaughn and colleagues (2013, 2015) utilized a randomized block study design with randomization occurring at student and class levels to test intervention efficacy in 8th grades. At first, students were randomly assigned to classes, then a treatment or control condition was randomly assigned to teacher's classes, with teachers having an odd number of classes ending up with an additional treatment condition class (i.e., 5 classes = 3 treatment + 2 comparison). Eighth grade students assigned to treatment classes received intervention during their social studies classes via three 10-day cycles that were aligned with three distinct units (Colonial America, the Road to Revolution, and the Revolutionary War), resulting in 30 sessions of intervention over the course of six to eight weeks. Students assigned to the control condition received business as usual instruction that covered the same curricular material intended to improve content knowledge and reading comprehension over the same period of time. Intervention differed from

control in terms of the delivery of the content and not in the curriculum provided. For detailed information about the intervention readers are referred to Vaughn and colleagues (2013, 2015).

Measures

Standardized reading comprehension measure. The Gates-MacGinitie Reading Tests - Reading Comprehension (GMRT-RC; MacGinitie, MacGinitie, Maria, & Dryer, 2000) is a group-administered, standardized measure of general reading comprehension. Students read passages and answer multiple-choice questions related to the passage. Items test recall of information and/or inferences based on the texts. Forms S and T of the test for grade level 7/9 were administered to all students. Forms T and S were administered as pre- and posttest measures, respectively. Each test form is comprised of 11 passages and contains 48 items. The internal consistency reliability was equal to .89 or higher for each test form.

Researcher-developed reading comprehension measure. The Assessment of Social Studies Knowledge - Reading Comprehension (ASK-RC; Vaughn et al., 2013) is a researcher-developed measure of reading comprehension that includes 21, four-option, multiple-choice items. The assessment consists of 3 reading passages differing in both length and overall difficulty (Lexile range = 1090 - 1140; word count range = 312 - 349), each of which is related to content covered in the intervention in the three 10-day cycles. The intervention is focused on building comprehension through reading, not on the teaching of specific content. In the assessment, students read a passage silently and immediately afterwards answer 7 multiple-choice questions about the passage. The internal consistency reliability was equal to .85.

Text features. Text characteristics for the GMRT-RC and ASK-RC passages were evaluated using Coh-Metrix and the Lexile Analyzer. The average word frequency, average sentence length, narrativity, syntactic simplicity, word concreteness, referential cohesion, and

deep cohesion of the GMRT-RC (forms S and T) and ASK-RC passages were measured utilizing the Coh-Metrix Text Easability Assessor (Graesser, McNamara, & Kulikowich, 2011). The Lexile level of the passage was obtained via the Lexile Analyzer which is available from Metametrics. For more detailed descriptions of these text features and their effects on reading comprehension, see Kulesz et al. (2016). Table 3 provides means and standard deviations for text features broken down by test type (the GMRT-RC forms S and T, and ASK-RC).

[INSERT TABLE 3 HERE]

Data Analytic Approach

A two-stage data analytic approach was implemented to examine the role of text features on item difficulty. Although the bulk of the modeling could have been carried out using joint estimation of item difficulties and the effects of text characteristics on item difficulty, we adopted a two stage estimation approach for ease of implementation and estimation. If we had elected to focus on reader-text interactions affecting item difficulty, joint modeling would have been required. Explicit testing of interaction effects with intervention would require joint modeling, and would have been pursued if we had evidence of differential effects of intervention on ASK and GMRT. The first step in our two-stage analysis involved alignment of the ASK-RC and GMRT-RC assessments (as described above) to obtain item difficulties that were used in subsequent explanatory models. The second stage of data analysis focused on describing similarities and differences between the ASK and GMRT test forms and examining the effects of text features on item difficulties.

Test alignment. The first stage of the analysis involved placing all items on a common scale. We used MPlus to fit a confirmatory factor model for all items from forms S and T of the GMRT-RC and the ASK-RC at the pre-test and post-test. This model constrained the ASK-RC

items to operate invariantly between the pre-test and post-test in order to place all items on a common scale. Because the study design confounded the GMRT-RC form with the timing of administration (form T at pre-test and form S at post-test) it was not possible to constrain the GMRT-RC items equal across pre- and post-test. Subsequent to placing all items on a common scale, we used a linear transformation[1] to rescale the item difficulties onto the scale of the GMRT Extended Scale Score, which is a developmental scale for reading comprehension similar to the Lexile scale.

Comparing Tests and the Effects of Text Features on Item Difficulties. The second stage of data analysis focused on: (1) examining correlations between texts features within tests (ASK-RC and GMRT-RC), (2) predictive discriminant analysis to examine differences in the test construction in terms of text features, and (3) regression models to examine the effects on text features on the item difficulties estimated in stage 1 analyses. Regression models included fixed effects of test type, text feature, and their interaction. Interactions were first examined individually in order to mitigate the effects of collinearly, and then were examined collectively to address possible redundancy across the separate models. Interactions between text features and test type were included to examine whether the effects of text features are generalizable across assessments or are test-specific. In other words, to examine whether differences in test construction influence the effect of text features on item difficulty.

Results

Test Alignment

Prior to fitting the confirmatory factor model to scale items, we first ensured that the ASK-RC and GMRT-RC items defined a common construct by examining the eigenvalues for

[1] We standardized the item difficulties to the mean and standard deviation of the GMRT ESS in our sample (Mean = 547; sd = 37) in order to scale all item difficulties to the GMRT ESS.

the pre-test items (i.e., ASK-RC at pre-test and GMRT-RC Form T items) separately from the post-test items (i.e., GMRT-RC Form S and ASK-RC items at the post-test). These analyses showed a very large first eigenvalue (24.2 at the pre-test and 24.8 for the post-test, accounting for approximately 35%-36% of the variance in the items). Although there were additional eigenvalues exceeding 1.0, the second eigenvalues were 3.4 and 3.8 at the pre-test and post-test, respectively, accounting for roughly 5% of the variance. These additional factors were not broken out by GMRT-RC versus ASK-RC items, but rather showed a variety of GMRT-RC and ASK-RC items loading on each minor factor.

In previous confirmatory factor modeling with the ASK-RC and GMRT-RC (Vaughn et al., 2013, 2015), the two tests were modeled separately and found to demonstrate measurement invariance between treatment and control groups, as well as over time for the ASK-RC. Because the GMRT-RC form was not the same at the pre-test and post-test, the GMRT-RC was not constrained invariant across time in those studies. Our study combines the two samples from these studies to align the GMRT-RC and ASK-RC. Given that the eigenvalues indicated a strong common factor, we proceeded to fit a model with a single pre-test and post-test factor using the KNOWN CLASS option in MPlus, with Treatment and Control group defining the known classes.

For the current study, we further constrained the model to correspond to the Rasch item-response model, constraining factor loadings equal across items and allowing items to differ from one another only in terms of the item thresholds. Thresholds for the ASK-RC were constrained equal for the same item between the pre-test and post-test. It is worth noting that the GMRT-RC was developed according to the Rasch-IRT model, providing additional support for imposing the Rasch-model constraints in the current study. We did consider allowing the

discrimination parameter to differ for GMRT-RC and ASK-RC items loaded on the same factor[2]. Although this model provided an improved "fit", it is conceptually less appealing and item thresholds were negligibly different (r = .99; average difference = 0.004). When item thresholds were transformed into Rasch model item difficulties (Asparouhov & Muthen, 2020), the difficulty parameter estimates also correlated .99 both within-test and across all tests, and differed by .03 on average. Thus, we proceeded with the Rasch parameter estimates in subsequent explanatory analyses[3] after rescaling them onto the GMRT Extended Scale. This rescaling allows us to interpret the parameters of the explanatory models in terms of the ability scale measured by the GMRT. Figure 1 presents the rescaled difficulties for each item plotted against the percent correct at the pre-test and post-test, along with some basic descriptive

[2] We also considered the two-parameter logistic model (2PL). The 2PL model allows the item discrimination parameter to vary across items. This parameter indexes the relationship between the item and the construct being measured. It is worth noting that if unidimensionality holds, then item discriminations must be uniform. In our case, the 2PL provides a statistically better fit to the data. However, ability estimates from the Rasch model and the 2PL correlated .99025 for the Pre-Test and .99163 for the Post-Test. Clearly, the models provide a difference without a distinction in terms of estimating ability. When we examined the effect of the 1PL vs 2PL on item difficulties, the correlation between item difficulties is .91. However, we observe something interesting within test. Specifically, the correlation for ASK is .87, whereas for GMRT the correlation is .92 and .95 for Form S and T, respectively. The low correlation for ASK stems from one very difficult item whose 2PL difficulty falls far from the difficulty of all other items. Omitting that item, the Ask correlation is .98 and the overall correlation among the difficulty estimates is .95. Thus, letting the item discriminations differ across items significantly affects a single item difficulty for one very difficult ASK item (approximately 20% passing), making it even more difficult. Under the 1PL model or the 2PL model, the item is the most difficult ASK item. If we allow the discrimination parameters to differ across items, it simply becomes the most difficult item instead of the second or third most difficult item. It also becomes an extreme outlier in the distribution of item difficulties, which would give it undue influence in regression models of difficulty. For these reasons, we have concentrated on item difficulties from the 1PL model.

[3] Technically, the Rasch model assumes item responses are independent conditional on ability, an assumption that is often violated for tests of reading comprehension that use passages, because items within the same passage may covary conditional on ability. This problem can be addressed by creating testlets that sum the number correct across items associated with a specific passage. These testlets are then analyzed as polytomous ordered categorical test items and the number of items is reduced to the number of testlets. Models for the testlets can be formulated to estimate the difficulty level for the testlet, but not the individual items within the testlet. The other model parameters apply to score shifts from 0, to 1, to 2, etc., but individual item difficulties are lost. Thus, changing our estimation model to a testlets approach comes at a significant cost to our ability to relate item and text features to estimates of item difficulty and to compare the tests on the basis of these relations. Thus, we have estimated item difficulties under the Rasch model, but treat them as nested within paragraphs for modeling text effects on item difficulty. We did estimate item difficulties for the pre-test occasion using the 1PL model under PROC GLIMMIX in SAS treating items as nested within paragraphs/passages and ignoring nesting. Estimates correlated over .99 under these two approaches and also correlated over .99 with the Rasch difficulties from MPLUS.

statistics for item difficulties and the percent passing. The scaled score associated with an item shows the level of ability needed to answer the item correctly with probability equal to .5.

Test Differences in Text Characteristics and Their Effects on Item Difficulty

Correlations and Predictive Discriminant Analysis. Figure 2 presents scatterplot matrices of text features by test type (left panel = ASK; right panel = GMRT). As is readily apparent in Figure 2, the magnitude of relations among text features differed substantially across the two test types. While the text features were highly correlated in the ASK-RC, text features were more independent across GMRT-RC test items (small to moderate). For the ASK-RC, narrativity and syntactic simplicity were moderately to highly correlated (magnitude from |.34| to |.93|) with other text features (viz., average word frequency, average sentence length, word concreteness, referential cohesion, and deep cohesion). Word frequency was least related to other text features, with the magnitude of relations ranging from |.05| to |.45|. For the GMRT-RC form S, narrativity was most highly correlated to other text features, with the magnitude of relations ranging from |.16| to |.86|. Correlations between narrativity and other text features were small to moderate on the GMRT-RC form T, ranging from |<.001| to |0.44|. Average word frequency was the least correlated with other features on the GMRT-RC form T (from |<.01| to |.31|), whereas form S correlations ranged from |.32| to |.59|.

To examine the extent to which these features differentiate test items from the different tests, we conducted both three-group (ASK-RC, GMRT Form S, and GMRT Form S), and two-group (ASK-RC vs GMRT-RC) discriminant analyses. Because of the very different variability in features across tests, we examined both linear and quadratic discriminant functions based on the features in Table 3. We used paragraph as the unit over which text features were measured for ASK items (n = 10), and used passage as the unit for GMRT-RC items (n = 11 for each of

Form S and T). Analyses were run at the item-level (total n = 117) as well as the paragraph/passage level (total n = 32) using both linear and quadratic classification functions.

Discriminant analyses showed that items on the three tests differed in their distribution of features, and this finding was supported when comparing ASK-RC to GMRT-RC, as well as when comparing GMRT Form S to GMRT Form T. Items from the three tests were clearly different from one another in terms of the average text features (Pillai's Trace = 0.877, F = 12.17, *df_num = 14, df_den=218,* $p < .0001$). All multivariate criteria were consistent with this inference. Examining the univariate statistics, measures of narrativity ($p < .001$), syntactic simplicity ($p < .001$), word concreteness ($p < .001$), referential cohesion ($p < .001$), and sentence length ($p < .025$) contributed significantly to the discrimination between the GMRT-RC and ASK-RC items. When items were classified using the leave-one out method (Lachenbruch & Mickey, 1968), 7 ASK-RC items were misclassified, 5 as GMRT Form S items and 2 as GMRT Form T items, while 7 Form S and 4 Form T items were misclassified as ASK-RC items. In addition, 8 Form S and 3 Form T items were misclassified to the alternate GMRT-RC form. Using a quadratic classification function (i.e., classification that allows the within group covariance to differ across groups) yielded 13 total classification errors, 4 GMRT Form S items classified as ASK – RC items, and 5 GMRT Form T items misclassified as GMRT Form S items. No Form S items were classified as Form T items when the quadratic classification function was applied.

Significance tests for the foregoing analyses ignore the nesting of items within passages, which leads to non-independence across observations. When discrimination and classification were based on analysis of a single observation per paragraph/passage, the tests still differed in multivariate space (Pillai's Trace = .857, $F = 2.57$, with *df_num* =14, and *df_den*=48, $p < .0077$).

Again, all multivariate test criteria were statistically significant, but in this instance tests differed only on narrativity ($p < .002$), and to a lesser extent on syntactic simplicity ($p < .075$), with all other univariate p values above .12. The leave one out method misclassified 4 ASK-RC passages (3 as Form S and 1 as Form T), 4 Form S and 2 Form T passages as ASK-RC passages, and 2 Form S and 3 Form T passages were misclassified as belong to the other GMRT-RC form.

Regression models. The role of text features in explaining item difficulty was examined using multi-level regression with estimated item difficulty expressed on the GMRT-RC Extended Scale Score scale as the outcome and item and text features as predictors. Items were considered clustered within paragraphs for ASK-RC and passages for GMRT-RC. The ASK-RC items at the post-test were excluded from the analysis, as these are identical to the ASK-RC items at the pre-test. We examined each feature on its own, along with test-type (ASK-RC or GMRT-RC) and item type (Text Memory or Text Inference). We examined individual predictor models because of the correlations among the text features and the relatively small number of paragraphs/passages available to help isolate unique effects.

Variance components for the random intercepts model showed that roughly 55% of the variability in item difficulty resided at the paragraph/passage level (tau_00 = 558.91; residual = 454.84; ICC = 0.551). Item type was not statistically significant when entered alone ($p < .48$), but item difficulty was slightly higher on average for the ASK-RC, than for GMRT-RC (β = 20.5, s.e.=10.2, $p < .047$). Of the passage features, only narrativity ((β = -9.7, s.e.=4.49, $p <$ .0333) was significant when entered alone. However, when all features were examined together both narrativity (β = -12.4, s.e.=5.84, $p < .038$) and referential cohesion (β = 13.4, s.e.=5.21, $p <$ .012) accounted for some of the variability in item difficulty. Inclusion of the text features reduced variability in item difficulty by 32% (tau_00 = 381.53, s.e. = 153.10). Item type and test

were not significant when text features were included in the model. We also fit a reduced model that included narrativity, referential cohesion, and test to assess if the differences between tests was accounted for by the text features. This reduced model showed significant effects for narrativity (β = -13.3, s.e.=5.7, $p < .021$) and referential cohesion (β = 12.2, s.e.=4.4, $p < .008$), but not for test (β = 4.74, s.e.=11.9, $p < .691$), indicating that the difference in item difficulty across the two test types was attributable to differences in narrativity and referential cohesion. This reduced model accounted for roughly the same percentage of variance in item difficulties across passages (tau_00 = 371.07, s.e. = 137.8). A similar reduced model for item type produced similar results, with statistically significant effects for narrativity and referential cohesion, but not item type.

We also examined the possibility that the association of text features with item difficulty was moderated by test type, but found no evidence of significant interaction between text features and test type when multilevel regression was used. When clustering was ignored, there was evidence that narrativity interacted with test type such that items from more narrative passages tended to be more difficult for the ASK-RC test, whereas the opposite was the case for the GMRT-RC. However, a close examination of this interaction (see Figure 3) suggests that the interaction is driven in part by the narrow range of narrativity measures for the ASK-RC. All ASK-RC paragraphs tend to be low in narrativity and yet show a range of item difficulty with the most difficult items tied to the more narrative passages. In contrast, the GMRT forms show a wide range of narrativity, with a modest negative relationship between narrativity and item difficulty.

Sensitivity to Treatment Effects

We have shown that both the ASK-RC and GMRT-RC assess a common construct and that it is possible to place items from both tests on a common scale, making it possible to consider each test's sensitivity to treatment effects. To do so, we examined the number of items on each test that a student would be expected to get correct as a function of the magnitude of treatment effects. The number correct score is a sufficient statistic for estimating ability on a Rasch based test, but even for non-Rasch based tests, it is possible to estimate the number of items an individual of a given ability would be expected to get correct if they were to take the test[4]. We extended this idea to examine the number of items that an individual would be expected to get right if they took the ASK-RC or the GMRT-RC, either form S or form T, at the post-test given a treatment effect ranging from 0.1 to 0.5 standard deviations.

Figure 4 was created to visualize the roles of test, item difficulty, pre-test ability and treatment efficacy on the expected number of items answered correctly on the post-test. The top panel of Figure 4 plots the expected score at different ability levels for effect sizes from 0.1 to 0.5 SD units, whereas the bottom panel shows the difference in the expected number correct at different points on the ability scale for these same effect sizes. From the bottom panel of Figure 4 it is clear that an effect size of 0.5 would yield fewer than two additional items answered correctly on the ASK-RC for a low-ability student (e.g., one at ESS = 475, about 1.9 SDs below the mean), whereas students at that ability level would be expected to see a change of approximately 4 items on Form T of the GMRT-RC and roughly 8 items on GMRT-RC form S. At an ability level of 500, the ASK-RC has very low sensitivity because of the large gap in item

[4] The conditional independence assumption of the Rasch model facilitates our examination of this question, but it would be possible to carry out a similar analysis under different models for the probability of correctly answering a given item, or obtaining a specific score on a testlet of items. The analyses presented are consistent with the approach currently taken for scoring the GMRT-RC test forms, but more nuanced approaches are possible and may lead to different conclusions. Analyses based on the Generalized Partial Credit Model for testlets suggest that our present findings may understate the insensitivity of the tests to change for low ability students. However, firm conclusions in this regard warrant additional research, and should be subjected to experimental validation.

difficulties between the bottom three items and the next group of items. In contrast, Form T appears to be more sensitive to treatment effects for students at this ability level, at least as reflected by the change in the expected number of items answered correctly. All tests show high sensitivity just below the mean of 547, but very little sensitivity for high ability students, i.e., those more than 1.5 sd above the mean (ESS > 602.5).

While the raw score and change in raw score from pre-test to post-test is important, with standardized tests like the GMRT-RC, treatment comparisons are generally based on comparisons of scaled scores, which are transformations of the raw score. In the case of the GMRT-RC, the transformation of raw scores to extended scaled scores is different depending on the test form. In Figure 5 we present the raw score to scaled score conversion for forms S and T as overlapping curves in the top panel with the left hand curve showing the relationship in the Fall of Grade 8 and the right hand curve showing the relationship in the Spring. The curves are identical for a given form in the fall and spring, but differ across forms. As can be seen in the top panel, the ESS for a given raw score is slightly higher for form S than for form T for raw scores below 12, and in contrast are slightly higher for form T than form S for raw scores between 12 and 32. Granted, the differences are not large.

Despite the near overlapping curves, it is important to consider what happens when someone takes one form of the test at the pre-test and another form at the post-test. What is the expected "scaled score" change for individuals with different pre-test ability if they experience treatment effects of a given size? For simplicity, we assume that all individuals experience a treatment effect equal to the average treatment effect size. In the bottom panel of Figure 5, we graph expected extended scaled scores at the post-test on forms S and T for individuals of differing pre-test abilities who experience a treatment effect ranging from 0.2 to 0.5. As in

Figure 4, the small "+" signs at the bottom of the figure signal the difficulty of the individual items on each test form. To determine the expected "scaled score" we determine the expected number of items answered correctly based on person ability (i.e., pre-test ability plus the treatment effect measured on the ability scale) and item difficulty given our model with all items placed on the same scale. The expected raw score is rounded to the nearest integer, in so far as all raw scores are integers, and convert this expected raw score to a scaled score using the raw score to scaled score conversions graphed in the top panel of Figure 4. Each plot contains a line for Form S and a line for Form T and a reference line that runs diagonally through the plot that signals when the post-test expected extended scale score equals the pre-test extended scale score. Points below the reference line indicate expected post-test scores that are below the pre-test score, and points above the line signify post-test scores that exceed the pre-test score.

What stands out in the bottom panel of Figure 5 is the separation between expected scores on Forms S and T and the relative insensitivity of the scaled score to changes in ability, especially for low ability students until the treatment effect reaches about 0.4 standard deviation units. For effect sizes of 0.2 and 0.3, students who are moderately below average in ability (i.e., about one standard deviation unit below average) at the pre-test are expected to obtain a scaled score below their pre-test score, and this problem would be exacerbated if the pre-test were administered using Form T and the post-test administered using Form S.

Discussion

Over-alignment of outcome measures to intervention content is a serious threat to inferences about treatment effects that is somewhat unique to education research, although other psychological research focused on learning may be similarly vulnerable. The problem is sufficiently serious in education research to have earned mention in the Standards Reference of

the What Works Clearinghouse, and to have been singled out by Slavin in his examination of factors complicating reviews in education (Slavin, 2008; Slavin & Madden, 2011). However, the other side of this problem is the general sense among intervention developers that standardized tests seem to lack sensitivity to treatment effects. This flip side of the problem has received scant attention in education research, and it is exceptionally rare that a standardized assessment is chosen for inclusion in an intervention study because it has been demonstrated to be sensitive to ability change in the range of ability that is the target of the intervention research. In this study we attempted to examine in greater depth the question of alignment and sensitivity of experimenter developed and standardized tests of reading comprehension and their sensitivity to treatment effects by examining these questions in the narrow context of a specific set of studies focused on reading in 8th grade Social Studies.

In the case of this specific Social Studies reading intervention, we examined the question of alignment by examining the dimensionality of the items from the standardized and experimenter developed reading assessments. This analysis found little evidence to infer that the items from the different tests might be measuring different constructs. Examination of eigen values for pre-test items and for post-test items showed very large first eigen values in both cases, and although second eigen values were non-negligible (exceeding 1.0), they accounted for a small amount of variance. Moreover, there was no indication that these additional dimensions were unique to either the items from the experimenter developed test, or the items from the standardized test at either the pre-test or post-test.

Examination of passage characteristics suggested that passage construction in the experimenter developed test and the standardized test differed systematically from one another, and especially in the relationships among text features. Specifically, features appeared to be

more highly correlated across passages for the experimenter developed test, which also tended to be lower in narrativity than the passages on the standardized test. These findings are interesting, but we cannot ignore the fact that the experimenter developed test consisted of only three passages and 10 total paragraphs, which we used as the sampling frame for examining passage features. In contrast, the standardized tests consisted of 11 passages each, most of which were single paragraph passages. Overall the sample of passages is relatively small for any given test type, and suggests caution in making inferences about differences in test construction.

That said, a number of analyses suggested that the experimenter developed test was closer in construction to Form S of the standardized test than to Form T, and that Forms S was about as different from Form T as it was from the experimenter developed test. The strongest evidence in support of this claim comes from the discriminant analysis for paragraphs/passages when clustering is accounted for by selecting only a single observation per passage. In this case, the distance between the ASK and Form S was 3.73, which was quite comparable to the distance of 3.31 between Forms S and T, whereas the distance between ASK and Form T was almost double at 6.25. When passages were classified, 4 of 10 ASK passages were misclassified, 6 of 11 form S passages were misclassified, and 5 of 11 Form T passages were misclassified. Importantly, three of the four classification errors for ASK were to Form S, and 1 of 4 was to Form T, whereas 4 of 6 classification errors of Form S passages were to ASK, and 2 were to Form T. Only 2 Form T passages were classified as ASK passages. Taken together, the pattern of results obtained across the various discriminant and classification analyses of passages and items supports the inference that the experimenter developed test is about as different from the standardized test as one form of the standardized test is from the other, which is significant considering that the two standardized forms are designed to be equivalent to one another.

Examination of item difficulties through multilevel regression revealed that roughly 55% of the variance in item difficulty resided at the paragraph/passage level and that the ASK items tended to be more difficult on average than items from the standardized test. Examination of item difficulties for Forms S and T showed that item difficulty is not equally distributed across the two forms. In fact, Form T has a batch of items toward the lower end of the difficulty scale whereas Form S has a gap between its easiest and second most difficult item. This difference in the distributions of item difficulty across Forms S and T is apparent in Figures 1 and 4, where item difficulties are plotted, but also in the top panel of Figure 5 which shows the relationship between raw scores and ESS for the two test forms. These trace lines show that scaled scores increase more rapidly as a function of increased raw scores for Form S at the low end of the scale. The greater increase in scaled scores per unit increase in raw scores at the bottom of the scale is an indication of greater spread in item difficulty across items at the low end of the scale. However, it also signifies a potential diminished sensitivity of the test to changes in ability at the lower end of the test.

We examined the sensitivity of each test to changes in ability by simulating treatment effects of varying magnitude for individuals with varying ability at the pre-test. These analyses revealed that the three tests varied in their ability to detect change at different points on the distribution of ability. Specifically, these analyses showed an insensitivity of the ASK for low ability students with pre-test abilities in the range from roughly 480 to 510 (about 1.8 to 1.0 standard deviations below average). Perhaps more importantly, these analyses highlighted differences in the two standardized test forms to changes in ability for low ability students. For effect sizes up to 0.3, the expected number of items correct changed by fewer than 3 items on Form S for students between -2 and -1.5 standard deviations below average at the pre-test. In

contrast, the raw score change on Form T for the same students ranges from 3 to 7 items for the same effects sizes. This relative insensitivity to small changes in ability (.1 to .3 s.d. units) remains even when we account for the raw score to scale score conversion. If these effects hold up on further examinations, they would argue against using Form S as a post-test measure in intervention studies focused on low ability students in Grade 8.

There are a number of potential concerns with our analysis that stem from the design of the studies incorporated into the analysis. In both studies Form S was used as the post-test and Form T as the pre-test. It would have been preferable if both tests had been used on both occasions of measurement and randomly spiraled throughout the sample. Having both tests randomly assigned to participants at the pre-test and post-test would have allowed for a stronger equating of item difficulties across the tests in our analyses. We relied on the ASK items to equate items across the three tests, and thus, our equating is dependent on invariance of the ASK items across the pre-test and post-test. While this assumption appears plausible (see analyses in Vaughn and colleagues, 2013), the current analyses would be strengthened if the design had involved spiraling[5] of both forms at the pre-test.

A second limitation of our analysis is that it uses the expected number correct to obtain the expected extended scaled score without taking into account errors of measurement. While our approach is valid and reflects the students' true score on the form, an actual experiment using the tests will be based on observed scores where scores are measured with error, and tests of treatment effects are based on average differences between treatment and control groups. It

[5] Spiraling is a term that describes a simple process for approximating random assignment of test forms within a blocked sample, such as when students are nested within classrooms and there is a desire to have roughly equivalent numbers of students completing each form in any given classroom. The different forms to be tested are arranged in an alternating fashion into a stack and then distributed systematically throughout a classroom beginning at the top of the stack until all forms in the stack are distributed, or all students have received a form.

would be possible to account for error using the standard error of prediction, which would account for some shrinkage back toward the mean. However, we believe the true score simulations are the proper first step for this investigation into test sensitivity.

A third limitation of our approach is that we have relied on the Rasch model and the assumption of conditional independence across items to obtain the expected extended scaled scores by computing the probability of correctly answering an item based on the Rasch model. Our decision to focus on the Rasch model was two-fold. First, it allowed us to retain item difficulties for examining the effects of text features on item difficulty across the different tests. Secondly, the Rasch model underlies the current scoring of the GMRT, which uses a raw score to scale score conversion and does not rely on testlet scores. However, to the extent that the model is not suitable for the tests, the actual probability of correctly answering an item will differ from the Rasch model probability, and the expected raw score that is associated with a specific change in ability from pre-test to post-test may deviate from our expected score calculations. We believe that more extensive investigation of this specific aspect of our study is critical to developing a more complete understanding of these tests' sensitivity to treatment effects for students with different baseline abilities.

The foregoing notwithstanding, if our findings are correct, they suggest that an experiment using Form S as the pre-test and Form T as the post-test would have greater sensitivity for small treatment effects for low ability students in Grade 8 than a study using Forms T and S in reverse order. This conjecture could be examined in the meta-analysis literature on reading comprehension interventions, but we are unclear if there are a sufficient number of studies using the GMRT in grade 8 in both orders to yield a worthwhile examination of this question. Of course, it is also possible to examine this question empirically in a true

experiment of treatment efficacy by spiraling forms S and T at both the pre-test and post-test in a future study.

The findings from our study also have implications for response to intervention (Fletcher, 2018) in practice, and the use of response to intervention for making identification decisions. We generally accept the premise that test sensitivity is uniform throughout the range of the test, but these results suggest test sensitivity to change may vary dramatically across the distribution of pre-test abilities. Is the response to intervention low because the test is insensitive to change of the magnitude experienced by this student given the pre-test ability of the student, whereas the same change experienced by a student at a different pre-test ability is within the sensitivity range of the test? Whenever there is interest in measuring change and comparing individuals on the magnitude of change in response to intervention, the question of test sensitivity must be considered if valid inferences are to follow.

Nothing in our study should be taken to suggest short comings or limitations of the Gates McGinitie tests for the purposes for which they were designed, namely that of measuring the reading achievement of a broad range of students across many grade levels and occasions of measurement. The GMRT is an outstanding collection of tests with strong psychometric properties that are well-suited for measuring reading across development within children. Our focus here is on the question of sensitivity to treatment effects of a small to moderate magnitude for low ability students over a narrow time frame. While these results for grade 8 suggest the test may not be sensitive to small effects for low ability students in grade 8, we have not examined this problem for students in grade 7 or 9, who would ostensibly be using the same test forms. However, the distribution of ability for these students would be different, and the pattern

of findings might vary, even if the analyses above for grade 8 are found to hold up on replication and cross-validation. Thus, it is important not to read more into these findings than is warranted.

We believe that these findings suggest that reading intervention researchers may wish to embark on more careful simulations when designing experiments to test treatment effects when treatment targets a specific range of student abilities. If static, on-grade level forms for standardized tests are found to be insensitive to treatment effects in the expected range, researchers may wish to consider standardized tests that use adaptive item selection. Such tests would be expected to show greater sensitivity to change provided that item pools are deep enough. Another possibility is to use the ESS, for tests that provide one, to examine the sensitivity of off-level forms to change in the expected range. For example, if we were designing an intervention for low ability students in grade 7, we would examine the sensitivity of the grade 6 form of the GMRT to intervention effects in the ability range of the target population in addition to examining form 7/9 that was examined in the current study. Designing intervention studies would be facilitated if test publishers could be convinced to share item calibration results with intervention researchers so that simulations of treatment studies could be conducted prior to taking interventions into the field. Without this information, the interventionist faced with a non-significant result may not be able to determine if the treatment was weak or the test was insensitive.

References Cited

Abedi, J. (2006). Psychometric issues in the ELL assessment and special education eligibility. *Teachers College Record*, *108*(11), 2282.

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, *29*, 137–164.

Boyle, Joseph. (1996). The Effects of a Cognitive Mapping Strategy on the Literal and Inferential Comprehension of Students with Mild Disabilities. *Learning Disability Quarterly. 19*. 86-98. 10.2307/1511250.

Cheung, A. C., & Slavin, R. E. (2015). How methodological features affect effect sizes in education. *Best Evidence Encyclopedia.*

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, *10*, 277–299.

Fletcher, J.M., Lyon, G.R., Fuchs, L.S., & Barnes, M.A. (2018). *Learning Disabilities: From Identification to Intervention*, *2nd Edition*. New York: Guilford.

Francis, D.J., Fletcher, J.M., Catts, H.W., & Tomblin, J.B. (2005). Dimensions affecting the assessment of reading comprehension. In S.G. Paris & S.A. Stahl (Eds.), *Children's reading comprehension and assessment,* (pp. 369–394) Mahwah, NJ: Lawrence Erlbaum.

Francis, D.J, Kulesz, P.A., Benoit, J.S. (2018). Extending the simple view of reading to account for variation within readers and across texts: The complete view of reading (CVR_i). *Remedial and Special Education, 39*(5)*,* 274-288.

Garcia, J. R., & Cain, K. (2014). Decoding and Reading Comprehension: A Meta-Analysis to Identify Which Reader and Assessment Characteristics Influence the Strength of the Relationship in English, *Review of Educational Research, 84(1)*, 74-111.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36(2)*, 193-202.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General, 111,* 228–238.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*, 281–300.

Keenan, J.M., & Meenan, C.E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities, 47(2)*, 125-135.

Kendeou, P., Papadopoulos, T. C., Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. Learning and Instruction, 22, 354-365.

Kulesz, P.A., Francis, D.J., Barnes, M.A., Fletcher, J.M. (2016). The influence of properties of the test and their interactions with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology. 108*(8), 1078-1097.

Lachenbruch, P.A. & Mickey, M.R. (1968) Estimation of Error Rates in Discriminant Analysis, *Technometrics, 10:1*, 1-11, DOI: 10.1080/00401706.1968.10490530

Linderholm, T., & van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of educational psychology*, *94*(4), 778-784.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch. W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14(1),* 1-43.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47(4),* 292-330.

Meyer, B. J., & Ray, M. N. (2011). Structure strategy interventions: Increasing reading comprehension of expository text. *International Electronic Journal of Elementary Education*, *4*(1), 127-152.

Mitchell, M. J., & Fox, B. J. (2001). The effects of computer software for developing phonological awareness in low-progress readers. *Literacy Research and Instruction*, *40*(4), 315-332.

Nation, K., & Snowling, M. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading*, 27, 342-356.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Schwartz, R. M. (2005). Literacy Learning of At-Risk First-Grade Students in the Reading Recovery Early Intervention. *Journal of Educational Psychology, 97*(2), 257–267. https://doi.org/10.1037/0022-0663.97.2.257

Slavin, R. E. (2008). What works? Issues in synthesizing education programs. *Educational Researcher*, *37*(1), 5–14.

Slavin, R. E., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*, 370–380.

Turner, J. E., Valentine, T., & Ellis, A. W. (1998). Contrasting effects of age of acquisition and word frequency on auditory and visual lexical decision. *Memory & Cognition, 26,* 1282–1291.

Vaughn, S., Roberts, G., Swanson, E. A., Wanzek, J., Fall, A. M., & Stillman-Spisak, S. J. (2015). Improving middle-school students' knowledge and comprehension in social studies: A replication. *Educational Psychology Review*, *27*(1), 31-50.

Vaughn, S., Swanson, E. A., Roberts, G., Wanzek, J., Stillman-Spisak, S. J., Solis, M., & Simmons, D. (2013). Improving reading comprehension and social studies knowledge in middle school. *Reading Research Quarterly, 48(1),* 77–93.

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.

Table 1

Sample Characteristics by Intervention Condition

Variable Name	Treatment	Control	Total
Gender, %			
Male	28.2	19.9	48.1
Female	29.1	21.3	50.4
Missing	0.8	0.6	1.4
Free Reduced, %			
Receives	18.0	13.2	31.2
Does not receive	29.4	21.0	50.4
Missing	10.8	7.5	18.3
LEP, %			
Currently in LEP	2.0	2.0	4.0
Not in LEP	55.0	39.0	94.0
Missing	1.0	0.6	1.6

Table 2

Means and Standard Devations for Pretest and Posttest Reading Comprehnesion Scores by Intervention Condition

Variable Name	Treatment			Control		
	N	*M*	*SD*	*N*	*M*	*SD*
Pretest GMRT-RC, T, ESS	1139	545.62	36.96	818	547.82	37.09
Pretest GMRT-RC, T, Raw Score	1139	32.46	10.22	818	32.94	10.24
Pretest ASK-RC, Total Correct	1139	29.12	11.70	818	29.52	11.23
Posttest GMRT-RC, S, ESS	1030	542.34	36.96	717	543.63	37.84
Posttest GMRT-RC, S, Raw Score	1030	31.84	10.14	717	32.15	10.23
Posttest ASK-RC, Total Correct	1021	38.37	13.96	712	35.67	12.63

Note. GMRT-RC = Gates Macginitie Reading Comprehension; ASK-RC = The Assessment of Social Studies Knowledge - Reading Comprehension; ESS = Extended Scale Score.

Table 3

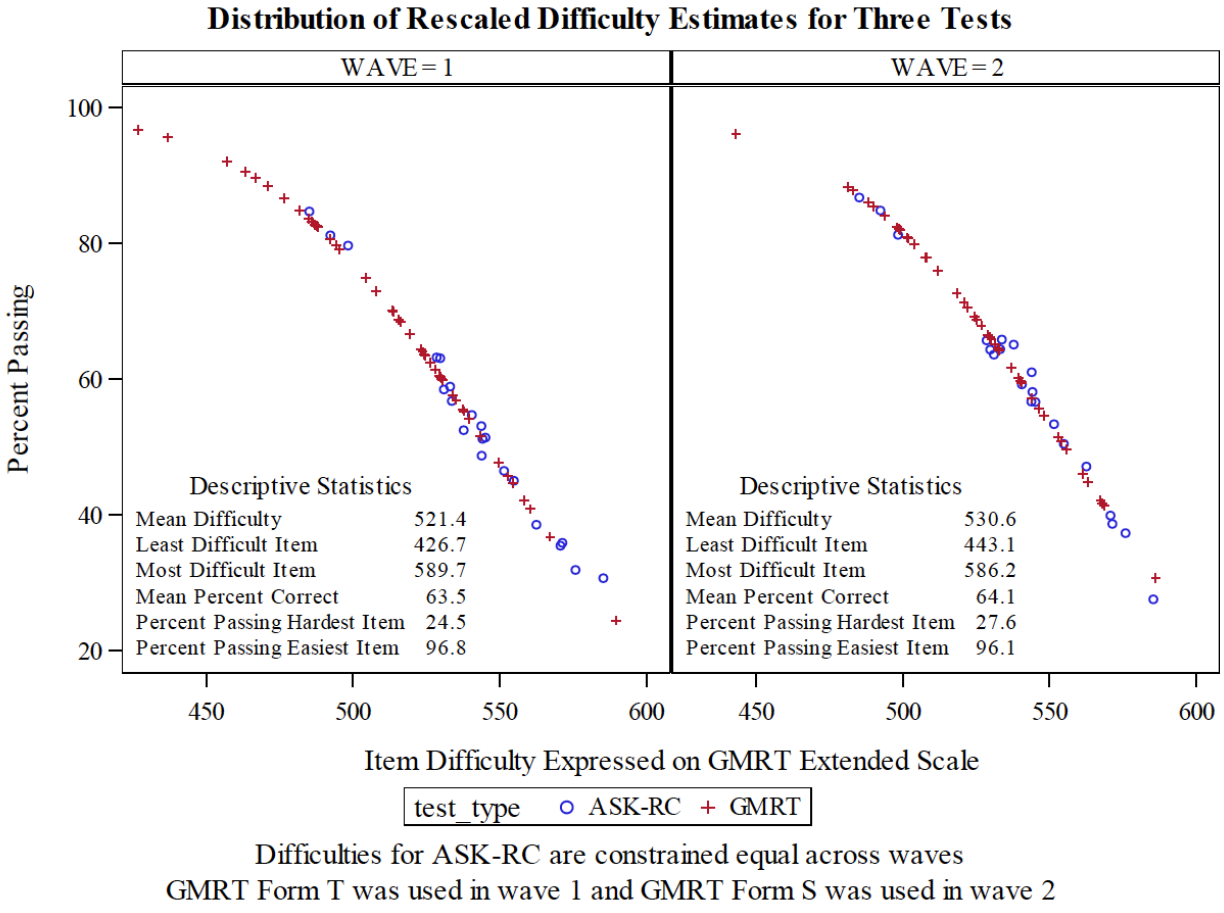
Means and Standard Deviations for Text features by Test Type

	ASK-RC *n of questions* = 21	GMRT-RC, S *n of questions* = 48	GMRT-RC, T *n of questions = 48*
Narrativity, *M (SD)**	-1.01 (0.19)	0.15 (0.82)	0.73 (0.84)
Syntactic Simplicity, *M (SD)**	0.66 (0.87)	-0.04 (.85)	-0.54 (0.92)
Word concreteness, *M (SD)**	-.57 (1.08)	0.24 (0.90)	0.26 (0.83)
Referential cohesion, *M (SD)*	-0.31 (0.91)	-0.27 (0.93)	0.54 (0.94)
Deep cohesion, *M (SD)*	-0.14 (0.71)	0.20 (1.20)	-0.08 (0.99)
Sentence length, *M (SD)**	-0.43 (0.77)	0.29 (0.88)	0.09 (1.17)
Word frequency, *M (SD)*	0.07 (0.39)	0.23 (1.50)	-0.08 (0.73)
Lexile range	1090L-1140L	800L-1400L	500L-1600L
Word Count range	312-349	79-170	57-157

Note. GMRT-RC = Gates Macginitie Reading Comprehension; ASK-RC = The Assessment of Social Studies Knowledge - Reading Comprehension; * = statistically significant differences between mean text feature scores on the ASK- RC vs GMRT-RC (averaged across forms S and T) based on predictive discriminant analysis. All measures have been standardized to a sample mean of 0 and standard deviation of 1.0, with the exception of Lexile value and Word Count.

Figure 1

Plot of Rescaled Item Difficulties Against the Percent Correct



Relationship between item difficulty rescaled to the GMRT Extended Scaled Score and the percent answering an item correct on ASK and GMRT in Fall (Wave 1) and Spring (Wave 2).

Figure 2

Scatterplot Matrices of Text Features for ASK-RC and GMRT-RC Forms S and T

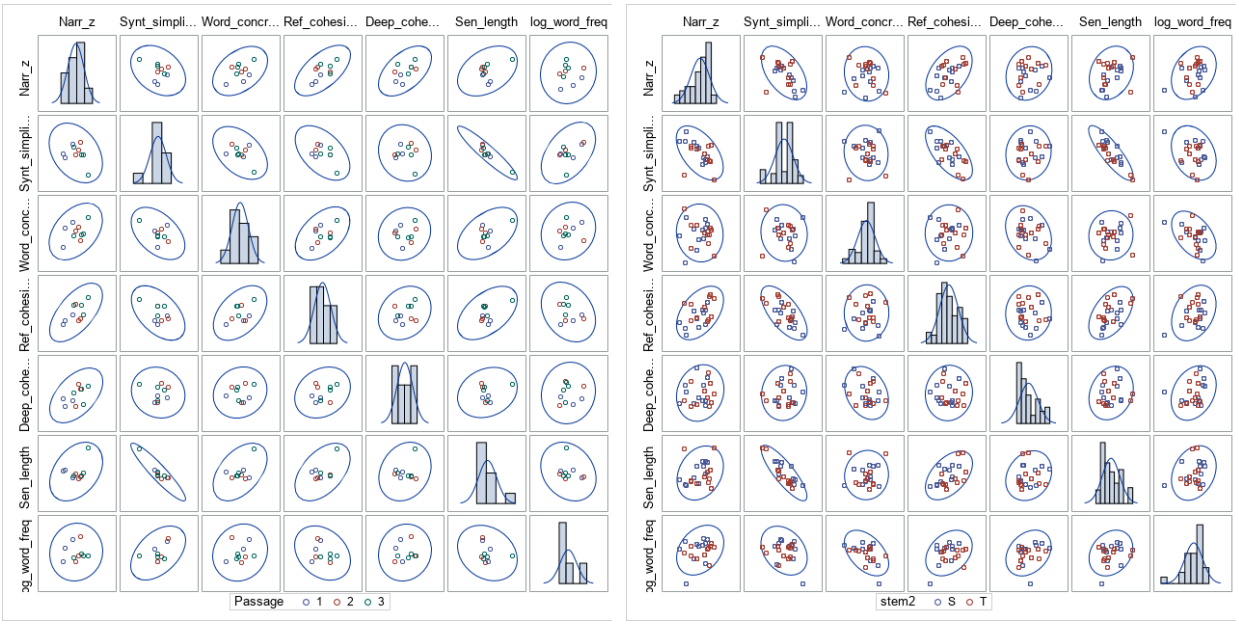
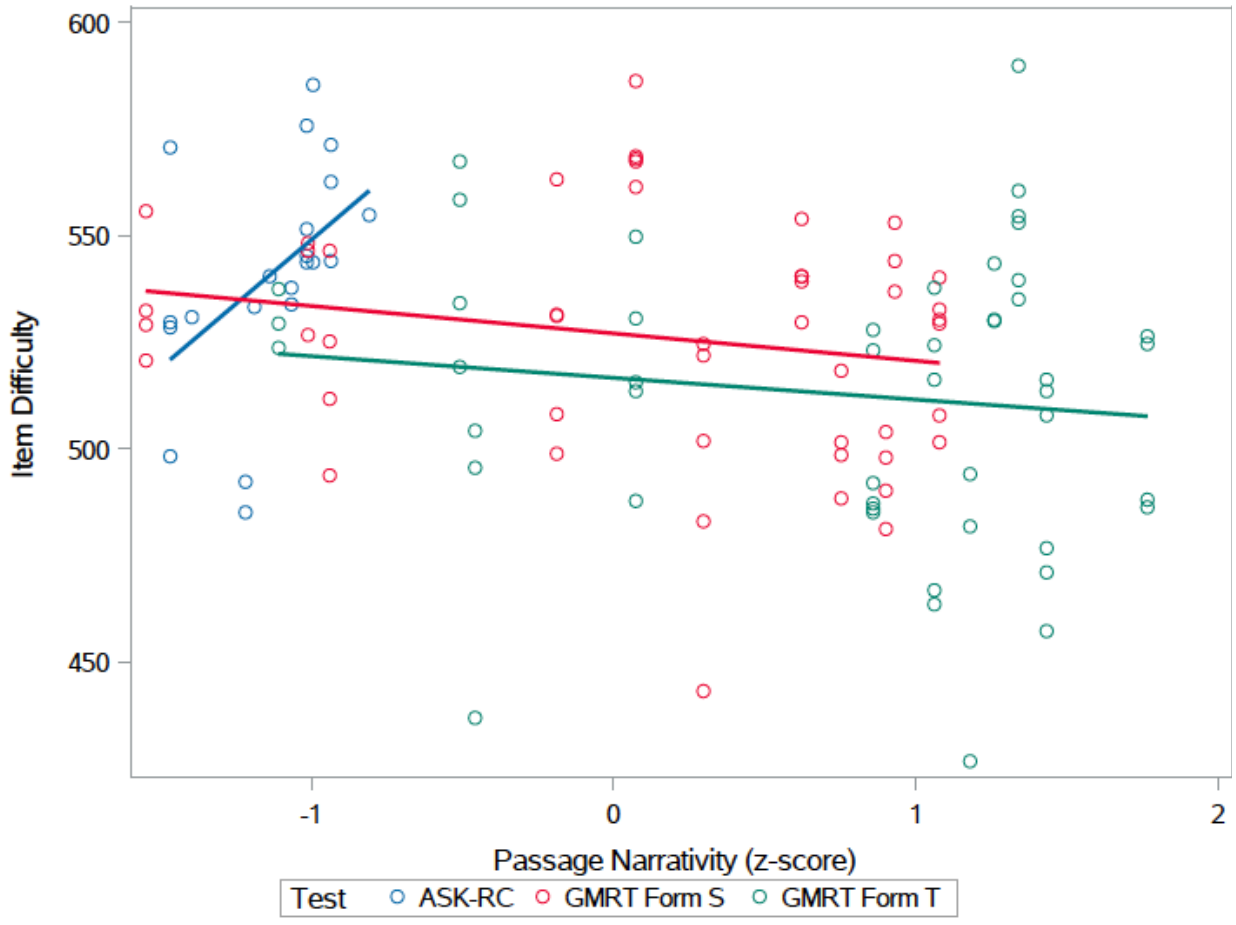


Figure 2. Scatterplot matrix of text features for the ASK-RC (left panel) and GMRT-RC (right panel). Please note that blue, red, and green colors on the left had side panel represent the ASK-RC passages. For the ASK-RC, we computed Coh-Metrix indices (text features) per paragraph within three ASK-RC passages since the ASK-RC paragraphs within any given passage were longer relative to the GMRT-RC paragraphs; (the GMRT-RC paragraphs were generally too short to yield meaningful indices). Blue and red colors on the right hand side panel represent the GMRT-RC test forms S and T, respectively. The text features are ordered from left to right and top to bottom in each panel as follows: Narrativity, Syntactic Simplicity, Word Concreteness, Referential Cohesion, Deep Cohesion, Sentence Length, and Log Word Frequency.

Figure 3

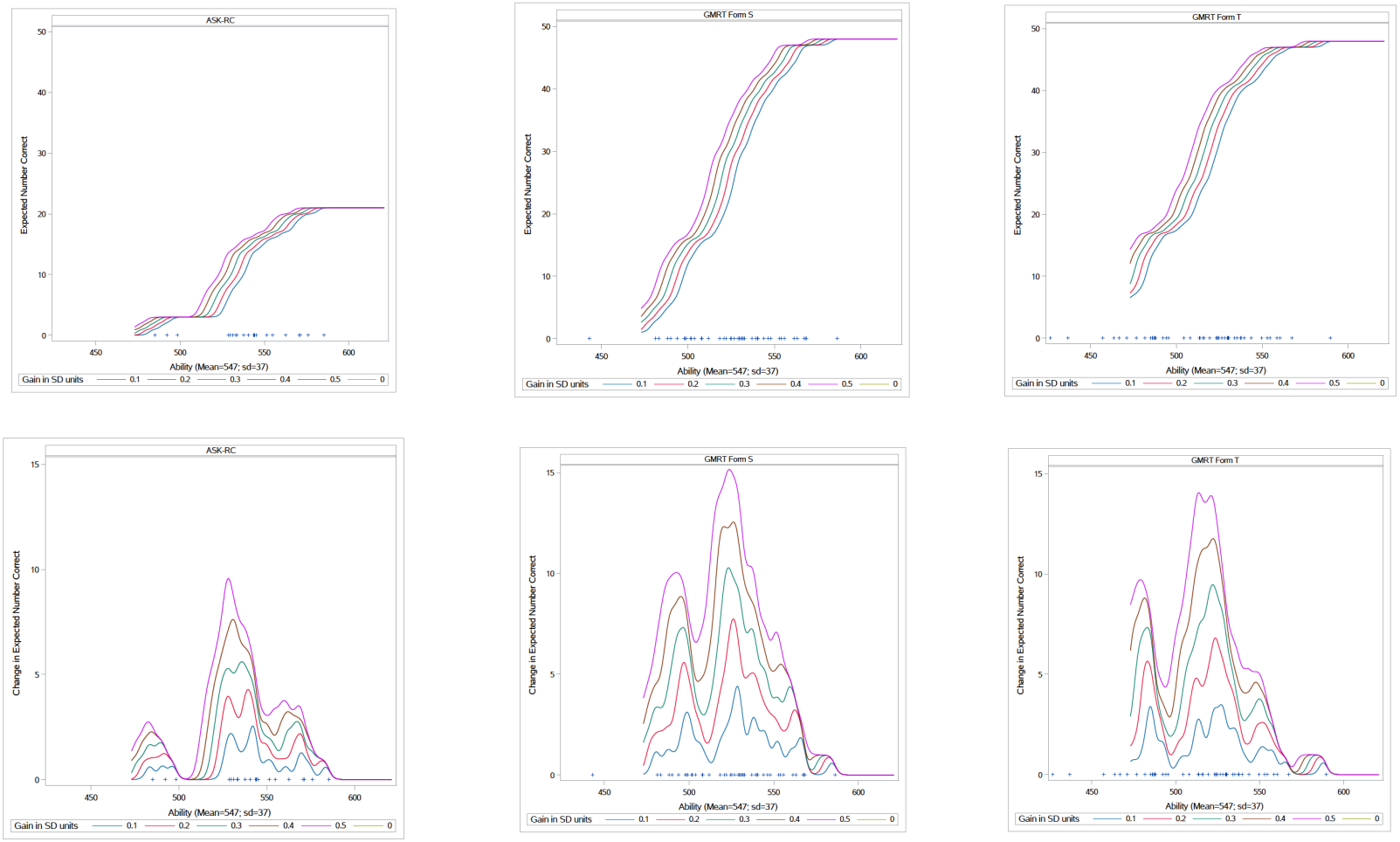
Interaction of Narrativity and Test Type on Item Difficulty



Plot of interaction between narrativity and test type on item difficulty with item difficult expressed on the GMRT Extended Scale Score scale. The plotted circles indicate values for individual test items. The plotted lines indicate the least squares regression line estimated within test.

Figure 4

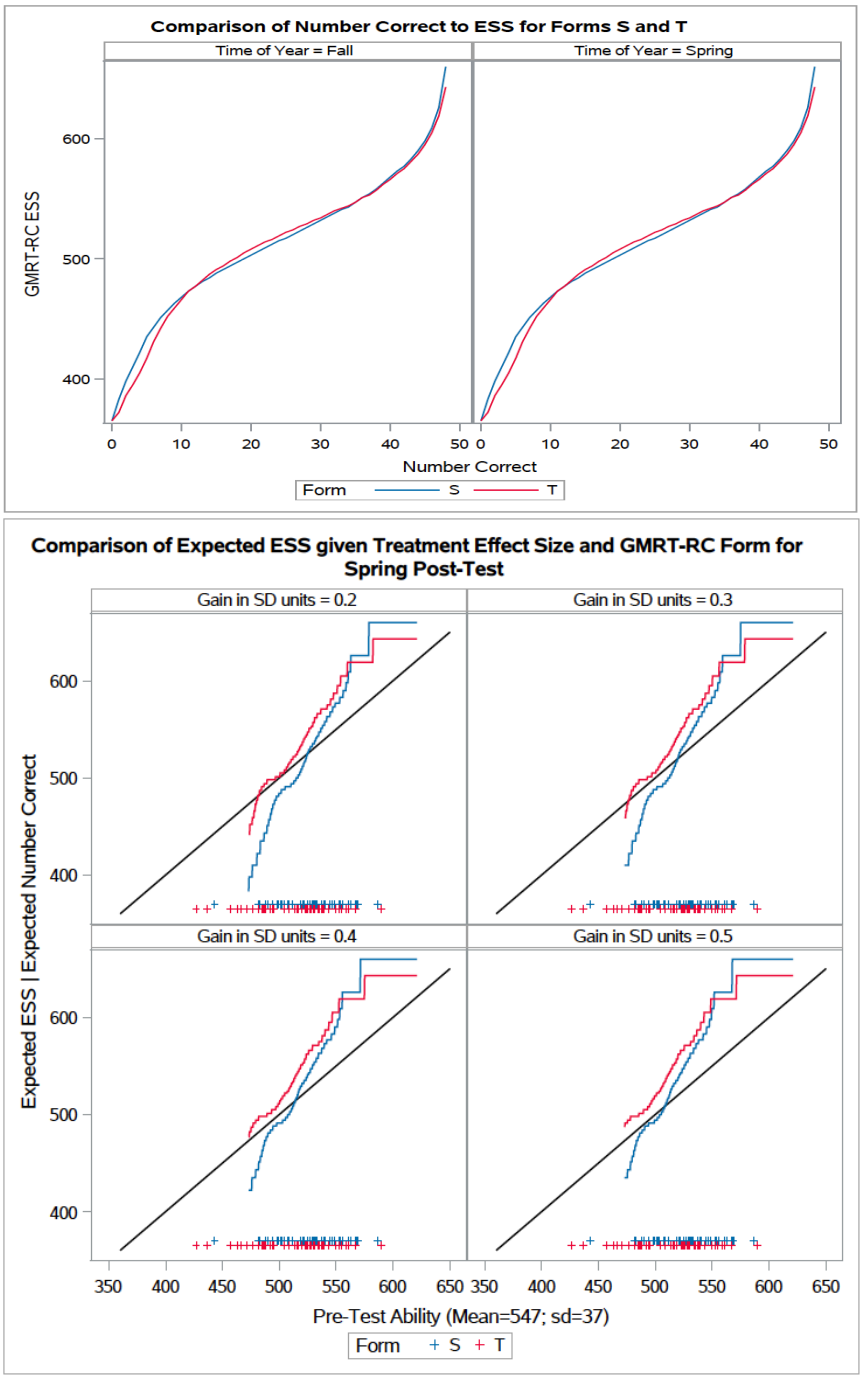
Panel plot of test sensitivity as measured by the expected number of items answered correctly as a function of pre-test ability, treatment efficacy, and test used to measure post-test ability.



Top Panel shows the expected number of items answered correctly if the test were used as a post-test measure plotted as a function of pre-test ability and a treatment effect of the specified magnitude from 0 to .5 sd units. The bottom panel shows the change in the expected number correct as a function of pre-test ability and the magnitude of the treatment effect. In both panels, the lines plot the value of the outcome for a specific gain in ability ranging from 0 to 0.5 sd units. The small + symbols at the bottom of each panel indicate the difficulty of individual items on the test. Item difficulties are plotted at the point on the ability scale where the probability of correctly answering the item conditional on ability is 0.5.

Figure 5

Raw score to scaled score conversions for GMRT-RC (top panel) and the expected scaled score after treatment for different effect sizes as a function of pre-test ability (bottom panels).



The top panel compares the conversion of raw scores to scaled scores for Forms S & T for Fall and Spring of Grade 8. The bottom panel compares the expected Extended Scale Score on Forms S and T in the spring of Grade 8 as a function of the size of the treatment effect (.2 to .5) and pre-test ability measured on the extended scaled score.