ETS GRE®

Assess Ability. Predict Performance.

ETS GRE® Board Research Report
ETS GRE®–16-02
ETS Research Report No. RR–16-20

Dimensionality Analyses of the *GRE®* revised General Test Verbal and Quantitative Measures

Frédéric Robin

Isaac Bejar

Longjuan Liang

Frank Rijmen

May 2016

The report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

****

****

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

GRE-ETS

PO Box 6000

Princeton, NJ 08541-6000

USA

To obtain more information about GRE programs and services, use one of the following:
Phone: 1-866-473-4373
(U.S., U.S. Territories*, and Canada)
1-609-771-7670
(all other locations)
Web site: www.gre.org

*America Samoa, Guam, Puerto Rico, and US Virgin Islands

GRE Board Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Dimensionality Analyses of the *GRE*® revised General Test Verbal and Quantitative Measures

Frédéric Robin,[1] Isaac Bejar,[1] Longjuan Liang,[1] & Frank Rijmen[2]

1 Educational Testing Service, Princeton, NJ

2 Association of American Medical Colleges, Washington, DC

Exploratory and confirmatory factor analyses of domestic data from the *GRE*® revised General Test, introduced in 2011, were conducted separately for the verbal (VBL) and quantitative (QNT) reasoning measures to evaluate the unidimensionality and local independence assumptions required by item response theory (IRT). Results based on data from the period immediately after the launch of the revised test and data from a year later showed that very little local item dependence was present and that a predominant single factor accounted for each of the data sets. These results provide evidence supporting the assumptions that underlie the use of the unidimensional 2-parameter IRT model for scoring and contribute evidence for the validity of GRE scores and their use.

Keywords *GRE*® General Test; dimensionality; factor analysis; bifactor model; multistage testing

doi:10.1002/ets2.12106

The purpose of this study is to evaluate the dimensionality of responses to the verbal (VBL) and quantitative (QNT) reasoning measures of the computerized multistage adaptive *GRE*® revised General Test[1] (henceforth referred to as GRE-MST), introduced in August, 2011. Such evaluation contributes to supporting a validity argument (Kane, 2006) for GRE scores and the continuing use by GRE of unidimensional item response theory (IRT) as the psychometric model.[2] Among the changes introduced by GRE-MST are the type of adaptive testing, the timing conditions, the item types, and the adoption of a two-parameter logistic (2PL) model, in contrast to the three-parameter logistic (3PL) model used by the previous GRE. The focus of this report is to document that the unidimensionality assumptions required by IRT holds in light of those changes.

The move to a multistage form of adaptive testing (Robin, Steffen, & Liang, 2014) allows test takers to skip items and return to them. In principle, this change makes it possible for test takers to "learn" from some items and then return to skipped items with a better chance of producing a correct response (Liu, Bridgeman, Gu, Xu, & Kong, 2015), which could introduce local dependencies among items. Other changes would seem to strengthen the unidimensionality assumption. The timing conditions are more generous under GRE-MST, which should reduce the presence of speededness and its negative effect on the use of IRT (Oshima, 1994). In addition, test takers are provided with an online calculator. Similarly, several content changes were introduced, especially changes to the VBL measure that would seem to contribute to unidimensionality. Most notably, VBL has evolved to exclude item types that emphasize decontextualized vocabulary, such as analogy and antonym items, which, as discussed below, have been the source of multidimensionality in the past. Currently, VBL consists of reading comprehension, text completion, and sentence equivalence items (Briel & Michel, 2014; Educational Testing Service [ETS], 2015a).

As noted, a further change introduced by the GRE-MST is the psychometric model used to estimate scores based on the observed responses to the test items (Robin et al., 2014). Even though the use of IRT was already in place in the previous version, GRE-MST relies on the 2PL model and the number of correct responses the test taker obtained on the test rather than the 3PL model and the pattern of item scores, which was used previously.[3] While estimating IRT theta (i.e., test taker's latent trait) from the raw number correct results in some loss of information (Thissen & Wainer, 2001; Yen, 1984a), advantages of this approach include the simplicity of the scoring process and its robustness to unexpected responses (Robin et al., 2014). As with any (unidimensional) IRT model, strong assumptions are made that responses to all items can be accounted for by a single ability and that there are no residual correlations among items dues to type,

Corresponding author: Frédéric Robin, E-mail: FRobin@ets.org

format, location, and so forth (local item independence). Now that the GRE-MST is operational and data are available, it is important to confirm that these assumptions are met by the VBL and QNT measures.

The repercussion of fitting a unidimensional model to multidimensional item response data has been studied, and there is consensus that unidimensional IRT models are robust to certain violations of unidimensionality (Gibbons, Immekus, & Bock, 2007). Nevertheless, one possible effect of departures from unidimensionality is to bias item parameter estimates, which would, in turn, bias the estimation of ability. Bias in the item parameter estimates could also have an indirect effect on the assembly of the test (Sireci, Thissen, & Wainer, 1991). In short, an evaluation of the unidimensionality and local independence assumptions remains necessary. We investigated these questions through intrameasure exploratory factor analyses (EFA) and the review of local dependence statistics, as well as through the application of multidimensional extensions of IRT applied to GRE-MST data at two points in time.

In the next sections, we review earlier studies concerned with the dimensionality of the GRE measures. We then describe the data and methods we used to assess the dimensionality of the GRE-MST and the results obtained. Finally we discuss the implications of our findings for GRE-MST and describe the need for further studies.

Review of Previous GRE Studies and Methodological Research

Although the GRE-MST is an extensive revision with respect to earlier versions of the test, it still intends to measure the same constructs as before. The GRE-MST also shares about half of the GRE computer adaptive test (GRE-CAT) VBL and most of the QNT item types,[4] which itself shared the same content as the restructured[5] paper-and-pencil test (GRE-PBT) that preceded it.[6] For these reasons, findings regarding dimensionality as well as the suitability of IRT to GRE data from studies conducted since the early 1980s with the GRE-PBT and the internal studies conducted in the early 1990s with the item-level adaptive GRE (GRE-CAT) are relevant to the GRE-MST. In this section, we briefly summarize the methods employed by these studies as well as their major findings.

The early factor analytic studies were primarily concerned with the dimensionality of the entire GRE, or intermeasure dimensionality, namely whether separate factors were needed for the three measures that comprised the GRE (VBL, QNT, and analytical [ANL]) as well as the dimensionality of each measure. These studies clearly established the need for separate dimensions across measures and support the GRE program caution (ETS, 2015b, p. 11) against combining scores across measures.

In summarizing studies prior to 1984, Stricker and Rock (1985) noted some instances of local dependencies or lack of unidimensionality:

> Some of these analyses also found factors defined by (a) difficulty of the item, (b) position of the item in the test section, (c) passages or diagrams common to more than one item, (d) regular mathematics items involving algebra, (e) regular mathematics items involving word problems (that is, verbally presented items with practical, concrete content), (f) data interpretation items entailing extraction of information, (g) data interpretation items entailing extraction and manipulation of information, and (h) quantitative comparison items for which the D alternative (that is, insufficient information) was correct. (p. 11)

Moreover, there was some evidence of intrameasure lack of unidimensionality. For example, Powers & Swinton (1981) reported findings based on EFA of interitem and item cluster tetrachoric correlations of two random samples of test takers' data from two GRE-PBT forms. Using an oblique rotation, two VBL (vocabulary and reading comprehension), one QNT, and one ANL (i.e., analytical) factor emerged. The design of the VBL measure then contained antonyms and analogies, which likely contributed to the presence of two verbal factors. As noted, the current GRE-MST does not include those item types.

In this report we are concerned only with intrameasure lack of unidimensionality that may be due to content or other factors. A later set of studies is more directly relevant to this report because they were concerned with the IRT assumptions and intrameasure dimensionality of GRE (PBT). For example, Dorans and Kingston (1985) evaluated the impact of departures from unidimensionality on equating results for the VBL measure and concluded that the impact was not severe, even though there was evidence of two dimensions for the VBL measure, which at the time included decontextualized vocabulary items (no longer in use), in addition to reading comprehension items. Similarly, Kingston and Dorans (1984) evaluated position or practice effects and noted that they were present primarily in the ANL section, which is no longer part of the GRE.

As a whole, these studies suggest that the dimensionality of the GRE General Test cannot be taken for granted and that it needs to be reevaluated as content, design, and conditions of administrations change. Based on the foregoing review, the nature of the changes introduced in GRE-MST, and the extensive development efforts that took place before the launch of the revised test (Briel & Michel, 2014), we hypothesize that the unidimensionality assumption will be met by both VBL and QNT. If violations of unidimensionality and local independence were to be present in GRE-MST, an important question is whether such departures matter in a practical sense. Methodological research bearing on the question suggests that fitting a unidimensional model to multidimensional data does not bias item parameter estimates, provided the secondary dimensions are small (e.g., Gibbons et al., 2007; Ip, Molenberghs, Chen, Goegebeur, & De Boeck, 2013; Reise, 2012). Nevertheless, it is important to explicitly evaluate the unidimensionality assumption with GRE-MST data.

Data

Analyses were conducted on VBL and QNT operational data collected a year apart. Two sets of VBL and QNT MST data were selected from data collected during the period we call the *jumpstart* period, lasting for the first 3 months after the August 1, 2011, launch of the GRE-MST. One set of VBL and QNT MST data was selected from data collected 1 year later, when the test program had fully transitioned to steady state operations.

All analyses were conducted using only domestic data because it accounts for approximately two thirds of the total population,[7] because the demographic composition of the domestic population has remained stable over time and because the operational IRT scale and calibrations are established using only domestic data (Robin et al., 2014).

Jumpstart Data

The jumpstart period was critical to the GRE program because it provided the data used (a) to calibrate the initial operational item bank from which the jumpstart and early steady state MSTs were assembled, and (b) to establish the new IRT and score reporting scales (Golub-Smith & Moses, 2014). The same operational MST design and specifications were used during the jumpstart and then in steady state.[8] However, the jumpstart MSTs were used for a longer period of time and incorporated the unscored sections[9] to augment the MST data collection and collect the minimum sample sizes required for accurate IRT calibrations and proper scaling (Robin et al., 2014). Furthermore, the GRE program offered a half-price retest incentive for tests taken in August and September to boost test-taker volumes and the representativeness of the jumpstart sample with respect to the total population. This incentive was needed because, despite the program's information campaigns about the new test and the publication of extensive test preparation materials ahead of the launch, test takers were generally apprehensive, as attested by the wave of postings on websites and social media advising test takers to take the older version while it was still available and avoid the new test as long as they could. This incentive also provided some compensation for jumpstart test takers who had to wait longer than the normal 2-week turnaround time to receive their official scores, starting November 1.

Figure 1 illustrates the jumpstart data collection design used to ensure accurate VBL and QNT MST calibration. Each test taker is assigned a VBL medium difficulty routing (R) section and, depending on performance, an easy (E), medium difficulty (M), or difficult (D) VBL section. In the same way, the test taker is assigned a QNT routing and an easy, medium or difficult QNT section. To augment the VBL or the QNT MST data, the unscored section is selected among the two MST sections not yet assigned, resulting in one of four assignment patterns: R-E-M, R-M-E, or R-M-D, or R-D-M. Thus, as most of the analyses we conducted focused on the MST medium difficulty forms,[10] we were able to use the R-M part of the R-D-M and the R-E-M data to augment the MST R-M data. Two jumpstart MSTs with some of the largest sample sizes of domestic test takers were chosen for this study: MST-1 and MST-2.

Steady State Data

As indicated earlier, the same two-stage MST design as described in Figure 1 has been implemented for ongoing steady state operations. In steady state, however, the unscored sections are used for their intended purpose: the pretesting and calibration of newly developed items. Thus, without the unscored sections to augment the data collection, traditional EFA were no longer possible; with each MST used for shorter periods of time to increase test security, smaller sample size datasets were available. One VBL and one QNT MST with the largest domestic sample size among the ones delivered in the higher volume months, in Fall 2012, were selected.

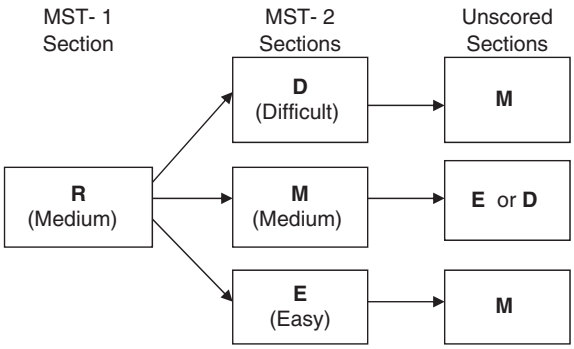

Figure 1 Jumpstart multistage (MST) design verbal (VBL) or quantitative (QNT). All test takers take a routing section (R) and are assigned a second-stage section, difficult (D), medium (M), or easy (E), based on cut scores or thresholds on the item response theory (IRT)-equated number-right score on R.

Table 1 Total Domestic Multistage Test (MST) Sample Size, Proportion of Test Takers Assigned to the Section, Mean, and Standard Deviation of Number Correct Scores Obtained With Each Operational MST Section

		Jumpstart				Steady state	
		VBL		QNT		VBL	QNT
Section		MST-1	MST-2	MST-1	MST-2	MST-1	MST-1
Routing	*N*	2089	2227	2167	1944	1276	1275
	Percent *N*	100	100	100	100	100	100
	M	10.6	10.1	10.1	10.4	9.9	10.2
	SD	3.9	4.4	4.1	4.9	4.9	4.9
Easy	Percent *N*	19	14	16	12	29	26
	M	10.9	10.4	10.5	8.8	10.3	10.3
	SD	3.7	3.7	3.4	3.2	3.7	3.5
Medium	Percent *N*	55	48	54	42	39	36
	M	12.7	10.8	11.8	8.5	11.5	11.5
	SD	3.6	3.7	3.6	3.0	3.3	3.5
Difficult	Percent *N*	26	40	29	46	32	38
	M	12.0	11.0	12.3	11.4	12.7	13.2
	SD	3.5	4.3	3.9	4.7	3.3	3.9

Note. VBL = verbal; QNT = quantitative.

Characteristics of the Data Sets Analyzed

Table 1 summarizes the sample sizes and number right performances obtained with each operational jumpstart and steady state MST investigated at the section level. All the number right means were within the range of 9 – 13 out of 20 possible, and the number right standard deviations were within 4 – 5 points for the routing sections assigned to all test takers and within 3 – 4 for the second stage sections. During the jumpstart period, the routing rates had been found to be somewhat unbalanced, with many MSTs having less than 20% of test takers routed to the lower difficulty form and more than 50% routed to the medium difficulty form, as it is the case here. In steady state, the routing thresholds were adjusted to produce rates closer to the intended 30%, 40%, and 30%, as we see here.

Table 2 summarizes the sample sizes and the scaled scores obtained with the augmented jumpstart and the steady state MST forms. As one can see, the sample sizes and, most importantly, the ranges of ability of the test takers assigned to the medium MST forms were larger with the augmented jumpstart MSTs: 6.7 to 8.4 versus 3.3 and 3.2.

Method

MST data are not well-suited for traditional EFA because the interitem tetrachoric correlations on which they rely cannot be estimated across the adaptive sections (Stage 2 sections in our case). While we hoped the jumpstart data augmentation could overcome this difficulty, preliminary analyses showed it was still not enough to produce meaningful results.

Table 2 Sample Size, Mean, and Standard Deviation of Scaled Scores Obtained With Each Augmented Jumpstart Multistage Test (MST) Form (As Described in Figure 1) and Each Operational Steady State Form

		Jumpstart				Steady state	
		VBL		QNT		VBL	QNT
Form		MST-1	MST-2	MST-1	MST-2	MST-1	MST-1
Easy	*N*	674	558	650	440	374	329
	M	147.4	145.5	143.1	138.0	144.2	141.5
	SD	5.8	5.0	5.1	4.4	4.2	3.7
Medium	*N*	2089	2277	2167	1944	500	459
	M	152.8	152.3	149.0	146.0	153.2	149.3
	SD	6.7	7.3	7.1	8.4	3.3	3.2
Difficult	*N*	836	1184	933	1094	402	485
	M	157.6	156.9	154.1	150.9	162.1	159.2
	SD	5.6	6.3	6.4	7.5	3.6	4.7

Note. VBL = verbal; QNT = quantitative.

Therefore, we focused the EFA analyses on the jumpstart medium forms. Because the medium forms do cover a large part of the item difficulty range and are assembled according to the same content specifications, we believe these analyses would provide very relevant information. Furthermore, as Table 2 shows, the sample sizes associated with the augmented jumpstart (MST-1 and MST-2) medium forms were above 2,000 and the standard deviation of ability values of 6.7 to 8.4 across VBL and QNT were representative of the total domestic values.

Local item dependence analyses were conducted on the same jumpstart medium forms using the Q3 statistic developed by Yen (1984b). The results of these analyses contributed to the evaluation of the appropriateness of using unidimensional IRT to score tests and to the interpretations of the factor analyses results.

Full information confirmatory factor analyses on the augmented jumpstart medium forms and the steady state medium forms and MSTs provided additional information for evaluating the dimensional structure of the VBL and QNT tests. Analyses of the MST (easy, medium, and difficult forms) datasets were possible because such methods make use of the full response vectors and avoid the limitations associated with estimating interitem correlations (Bock, Gibbons, & Muraki, 1988; Reise, 2012). Thus, despite the steady state ability restrictions and lower sample sizes (Table 2—total MST over 1,200 and medium form of 450), we were able to assess whether some changes in the dimensionality structure of the test took place between jumpstart and steady state, and whether the dimensionality of the test was consistent across MST easy, medium, and difficult forms. In the next sections, we describe the analyses conducted in more detail.

Exploratory Factor Analyses

For the EFA, tetrachoric correlation matrices were computed using the polycor package for the R statistical computing environment (Fox, 2010) and then analyzed using the CEFA[11] software. Cattell scree plots (Cattell, 1966) were used as rough indicators of the number of dimensions needed to account for the data. Rotated factor solutions were inspected to help identify the dimensional structure of the VBL and QNT measures.

Local Item Independence Analyses

Yen (1984b, 1993) described several factors including cheating, speededness, fatigue, practice, response format, dependence on a common stimulus, inadvertent hints from earlier items, and so forth that could cause departures from the local independence assumption used with unidimensional IRT. Some of these potential causes are typically prevented through judicious assessment design. But other potential causes that are inherent to test content, such as the inclusion of item sets that depend on a common stimulus,[12] may be more difficult to control. A number of statistics have been proposed to detect local item dependence (Chen & Thissen, 1997). However, Yen's (1984b) Q3 statistic probably remains the most widely used as it is simple and effective. Given a sample of test takers, Q3 is computed for item pairs as the correlation between item residuals (i.e., observed item score minus expected item score according to the model). In this study, we used the Q3 statistic to verify that local item dependence was not present, as well as to help interpret full information factor analyses if it was present.

Full Information Confirmatory Factor Analyses

Typically, factor loadings patterns obtained with educational tests are difficult to interpret because factors tend to be highly correlated and because items tend to load on multiple factors. However, as Reise (2012) has noted, "only recently has bifactor modeling been rediscovered as an effective approach to modeling *construct-relevant* multidimensionality" (p. 667). In this study, we used the bifactor analysis described by Gibbons and Hedeker (1992), extended by Rijmen (2009) and Jeon and Rijmen (2010), and implemented by Jeon, Rijmen, and Rabe-Hesketh (2014).[13]

As explicated by Rijmen (2010), there exists a formal relationship between the traditional IRT item parameters and item parameters of a multidimensional response model. The parameterization of an IRT model can be linearized by modeling the logit of a corresponding response probability. That is, for the 2PL, the probability of a correct response to the *i*th items is given by:

$$P\left(Y_i = 1\right) = \frac{\exp\left(a_i\left(\theta - b_i\right)\right)}{1 + \exp\left(a_i\left(\theta - b_i\right)\right)}$$

where a_i is the item discrimination, b_i is the item difficulty, and θ is the person ability parameter.

The equation above is equivalent to,

$$\text{Logit}\quad\left(P\left(Y_i = 1\right)\right) = \beta_i + \alpha_i\theta$$

where the intercept is

$$\beta_i = -a_i b_i$$

and the item loading is

$$\alpha_i = a_i.$$

With a bifactor model, the logit function is expanded according to a specific hypothesized dimensional structure expressed by

$$\text{Logit}\ \left(P\left(Y_i = 1\right)\right) = \beta_i + \alpha_{ig}\theta_g + \alpha_{ik}\theta_k$$

where g refers to a general factor, where all items load, and k refers to a specific subset of the items that define the kth secondary dimension. (Each dimension is defined by I_k items, with $\sum_{1}^{K} I_k = I$, the total test length.) The general factor loadings and intercepts are essentially equivalent to the corresponding unidimensional IRT model parameters and account for most of the variance in the data. The secondary factors account for some of the remaining variance to the extent possible given the orthogonality constraints that the model imposes on all the factors (resulting in all of them being uncorrelated). Once estimated, the fit of the unidimensional model and alternative bifactor models were evaluated to determine which model provided the best account of the data.

For that purpose, we used the Akaike information criterion (AIC)[14] and Bayesian information criterion (BIC)[15] information criteria (Akaike, 1973; Schwarz, 1978), which are typically provided as part of software outputs and are commonly used to identify the best fitting and most parsimonious model among alternatives (e.g., Ip et al., 2013; Rijmen, 2010). We also produced scatter plots of slopes (a_i versus α_{ig}) and intercepts ($-a_i b_i$ versus β_{ig}) to evaluate the extent to which the 2PL and bifactor general factor solutions were linearly related. In cases when a bifactor solution does provide a better fit to the data, then a linear relationship between the two sets of estimates would indicate that the test is "unidimensional enough" for the 2PL (Reise, 2012, p. 687).

Results

Jumpstart Exploratory Factor Analyses

Figure 2 shows the scree plots of the eigenvalues of the tetrachoric correlation matrices obtained with the VBL and QNT augmented datasets. It is recommended to pick the number of factors as the one that precedes the large drop (the "elbow" criterion). Although this criterion is subjective, it has been found to be helpful in practice (Hakstian, Rogers, & Cattell, 1982; Tucker, Koopman, & Linn, 1969). According to Figure 2, both VBL and QNT forms showed a very predominant main factor. By itself, the main factor accounted for 32.9% and 37.9% of the total variance for the QNT Forms 1 and 2, and 29.8% and 34.7% for the VBL Forms 1 and 2, respectively. To explore the possibility of higher number of factors, EFA

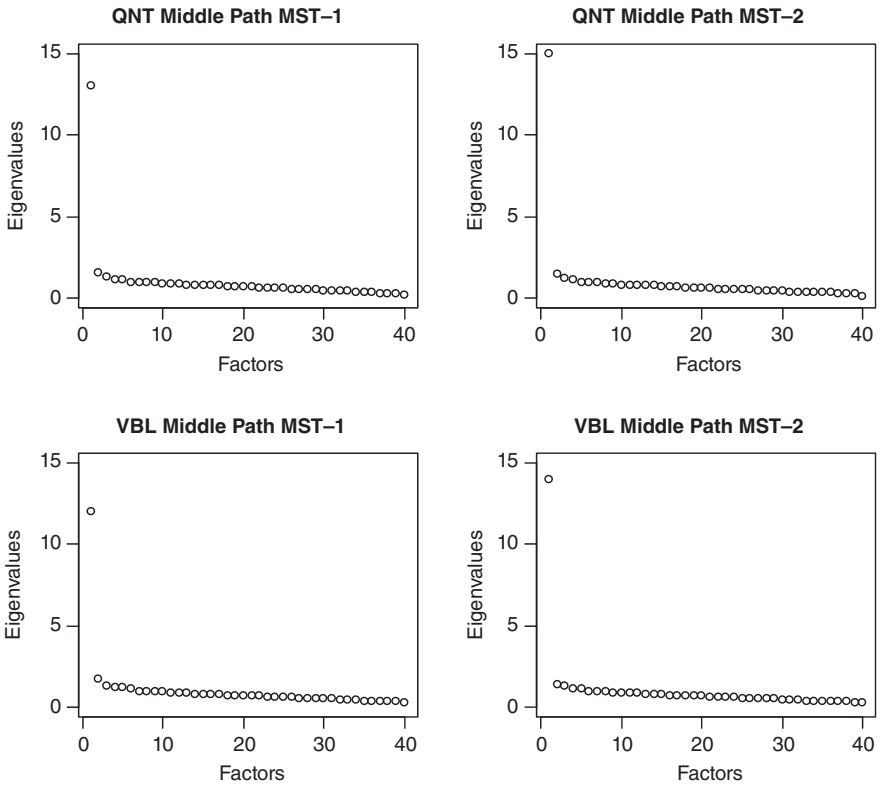

Figure 2 Eigenvalue scree plots in the multistage tests (MST) for the four jumpstart verbal (VBL) and quantitative (QNT) medium forms.

Table 3 Summary of the General Factor Loadings for the Four Jumpstart Verbal (VBL) and Quantitative (QNT) Medium Forms

	Minimum	25th percentile	Median	75th percentile	Maximum
QNT medium Form 1	0.31	0.47	0.56	0.63	0.83
QNT medium Form 2	0.40	0.50	0.61	0.68	0.90
VBL medium Form 1	0.09	0.43	0.52	0.61	0.74
VBL medium Form 2	0.27	0.50	0.56	0.65	0.78

using two, three, four, and five factors and oblique rotation[16] were conducted. However, none of the higher dimensionality models showed a clear and interpretable structure.

Table 3 provides a summary of the general factor loadings for the four medium forms. In all cases, nearly all the items loaded well on the general factor (median loadings above 0.5 and 25th percentile loadings above 0.4 for the four forms investigated).

The goodness of fit of the one-factor model was further evaluated using the root mean square error of approximation (RMSEA; Browne & Cudeck, 1992). As a general guideline, RMSEA < .05 indicates a close fit and .05 < RMSEA < .08 indicates a fair fit. The RMSEAs were 0.064 for both QNT forms, and 0.056 and 0.060 for the VBL Forms 1 and 2, respectively. These results indicate a fair model data fit for the one dimensional solution, but do not exclude the possibility of meaningful secondary dimensions.

Jumpstart Local Item Dependence Analyses

Although the EFA suggested that a single factor was sufficient to account for the QNT and VBL data, significant local item dependency could still exist, especially amongst the VBL reading comprehension and the QNT data interpretation item sets. Table 4 provides the empirical distributions of Q3 statistics as well as the expected and observed Q3 means for nonset- and set-based item pairs. As demonstrated in Yen (1993), the expected Q3 can be computed as $-1/(n-1)$,

Table 4 Number of Item Pairs, Expected Null Distribution Q3 Mean, Empirical Q3 Mean, and Empirical Q3 Distributions for Verbal (VBL) and Quantitative (QNT) Medium Forms

		Null Q3	Q3	Q3 distribution							
	N	mean	mean	[−.15,−.1)	[−.1,−.05)	[−.05,0)	[0,.05)	[.05,.1)	[.1,.15)	[.15,.2)	[.2,.25)
VBL medium Form 1											
All Item pairs	780	−0.03									
Non-set-based Items	765		−0.02		118	408	213	23	2	1	
Set-based Items	15		0.04			3	6	5	1		
set 1	6		0.00			3	3				
set 2	3		0.08				1	1	1		
set 3	1		0.07					1			
set 4	3		0.06					3			
set 5	1		0.04				1				
set 6	1		0.04				1				
VBL medium Form 2											
All Item pairs	780	−0.03									
Non-set-based Items	765		−0.01	1	81	428	244	11			
Set-based Items	15		0.05			4	4	6			1
set 1	6		0.00			3	2	1			
set 2	3		0.07				1	2			
set 3	1		−0.01			1					
set 4	3		0.07					3			
set 5	1		0.02				1				
set 6	1		0.22								1
QNT medium Form 1											
All Item pairs	780	−0.03									
Non-set-based Items	774		−0.02	1	100	446	213	11	3		
Set-based Items	6		0.03				5	1			
set 1	3		0.03				3				
set 2	3		0.03				2	1			
QNT medium Form 2											
All Item pairs	780	−0.03									
Non-set-based Items	774		−0.01		94	425	237	17	1		
Set-based Items	6		0.06			1	2	2	1		
set 1	3		0.03			1	1	1			
set 2	3		0.09				1	1	1		

Table 5 Unidimensional and Bifactor Model Data Fit for Verbal (VBL) and Quantitative (QNT) Medium Form 1

		Bifactor minus 2PL		
VBL	2PL	Item type	Guessing	Content
Number of parameters	80	120	120	120
AIC	96504	−326	−254	−138
BIC	96956	−101	−29	87
QNT		Content	Context	Format
Number of parameters	80	120	120	120
AIC	94260	−100	−178	−132
BIC	94714	128	50	96

Note. 2PL = two-parameter logistic; AIC = Akaike information criterion; Bayesian information criterion = BIC.

where n is the number of items. By design, item relationships (e.g., items providing clues to one another, sharing similar content or vocabulary, etc.) and speededness are controlled through the test assembly (Robin & Steffen, 2014). Thus, it was expected that the nonset-based medium form would provide a null Q3 distribution against which outliers exhibiting significant local item dependence could be identified.

As Table 4 shows, no nonset-based item pair Q3 statistic appeared as outlier. Only one VBL and no QNT set-based item pair was identified as an outlier. This confirmed that the IRT local independence assumption was warranted and confirmed the effectiveness of the GRE MST item development and automated test assembly in preventing related items (derived from the same parent, providing clues on one another, etc.) to be included in the same test.

Table 6 Verbal (VBL) Jumpstart Medium Form 1 Item Loadings for Three Alternative Bifactor Models

	Item type				Guessing			Content					
Items	Gen.	RC	TC	SE	Gen.	Yes	No	Gen.	Hum	SS	BS	PS	Oth.
Item 1	1.19		0.62		1.17	−0.23		1.15				−0.02	
Item 2	0.82		0.09		0.81	0.04		0.82	−0.10				
Item 3	1.24		0.42		1.23		0.32	1.22					0.10
Item 4	0.98		0.19		0.96		0.23	0.95					0.15
Item 5	0.95		0.16		0.96		−0.05	0.97		0.06			
Item 6	1.18		0.89		1.15		0.63	1.12	−0.38				
Item 7	0.62	0.31			0.65		−0.05	0.66	0.17				
Item 8	0.65	0.45			0.69	0.34		0.69			0.24		
Item 9	0.67	0.28			0.70	0.25		0.79	0.71				
Item 10	0.14	0.11			0.16	0.07		0.16	0.09				
Item 11	0.72	0.38			0.77	0.30		0.82	0.49				
Item 12	0.70	0.44			0.75	0.35		0.77	0.30				
Item 13	1.14			0.40	1.12	−0.32		1.07				0.05	
Item 14	1.07			0.60	1.02	−0.46		2.36					3.77
Item 15	1.35			0.87	1.39	−0.99		1.15	−0.45				
Item 16	0.98			−0.24	0.92	0.07		0.93				0.08	
Item 17	0.49	0.41			0.54	0.48		0.55					−0.12
Item 18	0.89	0.72			0.96	0.49		1.02				0.76	
Item 19	0.63	0.48			0.68	0.28		0.72				0.70	
Item 20	0.69	0.58			0.75	0.35		0.87				1.04	
Item 21	1.69		0.35		1.68	−0.26		1.69			−0.21		
Item 22	0.45		0.14		0.45	0.17		0.46		0.18			
Item 23	0.87		0.08		0.85		0.01	0.85					0.03
Item 24	1.11		0.11		1.10		0.13	1.11	−0.18				
Item 25	2.05		0.47		2.04		0.43	1.99	−0.06				
Item 26	1.74		0.26		1.71		0.29	1.77				−0.33	
Item 27	1.01	0.38			1.03	0.14		1.02		0.23			
Item 28	0.95	0.55			1.00	0.34		0.99		0.38			
Item 29	1.75	0.48			1.77	0.33		1.77		0.41			
Item 30	1.45	0.54			1.50	0.34		1.67		1.06			
Item 31	0.95	0.35			0.98	0.33		1.00		0.49			
Item 32	1.87			−0.43	1.69	−0.02		1.71	−0.21				
Item 33	1.44			−0.45	1.29	0.15		1.31			0.04		
Item 34	1.15			−0.30	1.05	−0.07		1.05		0.05			
Item 35	1.83			0.53	1.74	−0.40		1.66					0.23
Item 36	1.20	0.53			1.24	0.36		1.28			0.59		
Item 37	1.26	0.56			1.34		−0.38	1.30				0.26	
Item 38	1.06	0.87			1.14	0.68		1.09				0.38	
Item 39	1.33	0.69			1.38	0.45		1.34			0.22		
Item 40	1.23	0.65			1.42		−0.75	1.53			1.17		

Note. Gen = general factor; RC = reading comprehension; TC = text completion; SE = sentence equivalence; Hum = humanities; SS = social studies; BS = biological science; PS = physical science; Oth = other subjects. Item sets with more than one set member are shaded. When two sets are given next to each other, they are differentiated by shading intensity.

Jumpstart Confirmatory Bifactor Analyses

As suggested by past analyses of GRE and other similar tests, specific aspects of the constructs measured, item types, and formats (ETS, 2015a) could be related to the presence of meaningful secondary dimensions. Thus, without any clear indications of any other potential dimensional from the exploratory analyses, we focused on the following. To investigate this possibility, the verbal items were classified in three ways:

- *item type*, which includes reading comprehension (RC), text completion (TC), and sentence equivalence (SE);
- *guessing*, which includes two levels depending on whether the item is a traditional multiple-choice item (5-choice guessing) versus multiple selection multiple choice (MSMC) or numeric entry items, with which correct guessing is very unlikely; and

Table 7 Quantitative (QNT) Jumpstart Medium Form 1 Item Loadings for Three Alternative Bifactor Models

	Content					Context			Format				
Items	Gen.	Alg.	Ari.	DA	Geo.	Gen.	Real	Pure	Gen.	SSMC	QC	NE	MSMC
Item 1	1.48			−0.18		1.45	−0.05		1.43		0.30		
Item 2	0.58	0.12				0.58		0.17	0.57		0.14		
Item 3	1.20	0.13				1.20		0.05	1.20		0.05		
Item 4	1.03				0.44	1.02		0.06	1.02		0.06		
Item 5	0.84		0.22			0.87		0.53	0.84		0.54		
Item 6	1.67	1.93				1.19		0.64	1.14		0.62		
Item 7	1.21				0.11	1.22		0.02	1.21		0.09		
Item 8	1.20		0.53			1.19		0.39	1.15		0.29		
Item 9	0.67		−0.03			0.64	0.44		0.67			0.28	
Item 10	0.90	0.06				0.89		0.08	0.89	−0.01			
Item 11	1.15			0.11		1.14	0.24		1.15	0.15			
Item 12	0.90				0.48	0.89		0.08	0.90	−0.16			
Item 13	2.16	−0.31				2.19	1.02		2.24	0.82			
Item 14	0.84			0.50		0.82	0.57		0.86	0.68			
Item 15	0.79			1.01		0.71	0.57		0.74	0.67			
Item 16	0.84			0.37		0.83	0.28		0.84	0.36			
Item 17	1.53		0.27			1.52		0.04	1.54	−0.07			
Item 18	1.16			0.35		1.20		−0.28	1.18	0.44			
Item 19	0.90			0.06		0.90	0.04		0.93				−0.31
Item 20	1.62			0.18		1.62		0.02	1.70			−0.51	
Item 21	1.78	0.67				2.09		1.22	2.83		2.51		
Item 22	0.72			−0.06		0.72		0.05	0.71		0.18		
Item 23	0.97			0.24		0.96		0.09	0.96		0.17		
Item 24	1.14				0.71	1.09		0.29	1.07		0.24		
Item 25	0.95		0.06			0.93	0.30		0.94		0.12		
Item 26	1.37				0.20	1.38		−0.08	1.38		−0.04		
Item 27	0.83		0.46			0.80		0.14	0.80		0.05		
Item 28	1.41		0.08			1.38	0.50		1.50			0.76	
Item 29	1.43		0.48			1.39		0.30	1.39	−0.11			
Item 30	2.17		0.16			2.14	0.36		2.16	0.16			
Item 31	1.49	0.17				1.51		−0.15	1.50	0.37			
Item 32	2.09	0.18				2.11	−0.12		2.12	−0.17			
Item 33	1.26	−0.01				1.26		0.19	1.26	−0.05			
Item 34	1.61			0.41		1.59	0.62		1.60			0.29	
Item 35	1.16			0.26		1.16	0.51		1.16	0.31			
Item 36	1.19			0.28		1.19	0.41		1.19	0.31			
Item 37	1.66				0.34	1.69	−0.10		1.66	−0.05			
Item 38	1.49		−0.17			1.56		−0.44	1.48			0.03	
Item 39	1.39			0.34		1.50		−0.54	1.50				0.65
Item 40	2.54				0.54	2.51		−0.05	2.50	0.02			

Note. Gen = general factor; Alg = algebra; Ari = arithmetic; DA = data analysis; Geo = geometry; SSMC = single selection multiple choice; QC = quantitative comparison; NE = numeric entr; MSMC = multiple selection multiple choice. Item sets with more than one set member are shaded.

- *topic* including humanities, social science, biological science, physical science, and other subjects.

The quantitative items were classified in three ways:

- *content*, which includes four categories (i.e., algebra, arithmetic, data analysis, and geometry);
- *context*, which includes two levels depending on whether items are given a real life context or are presented in a pure mathematical way (referred to as *real* and *pure* in this report); and
- *format*, which includes single selection multiple choice (SSMC), quantitative comparison (QC), numeric entry (NE), and MSMC.

Three VBL and three QNT alternative bifactor models were fitted to the data using the above categorizations to postulate the secondary factors. For the 2PL and these alternative models, Table 5 provides the values of the AIC[17] and BIC[18]

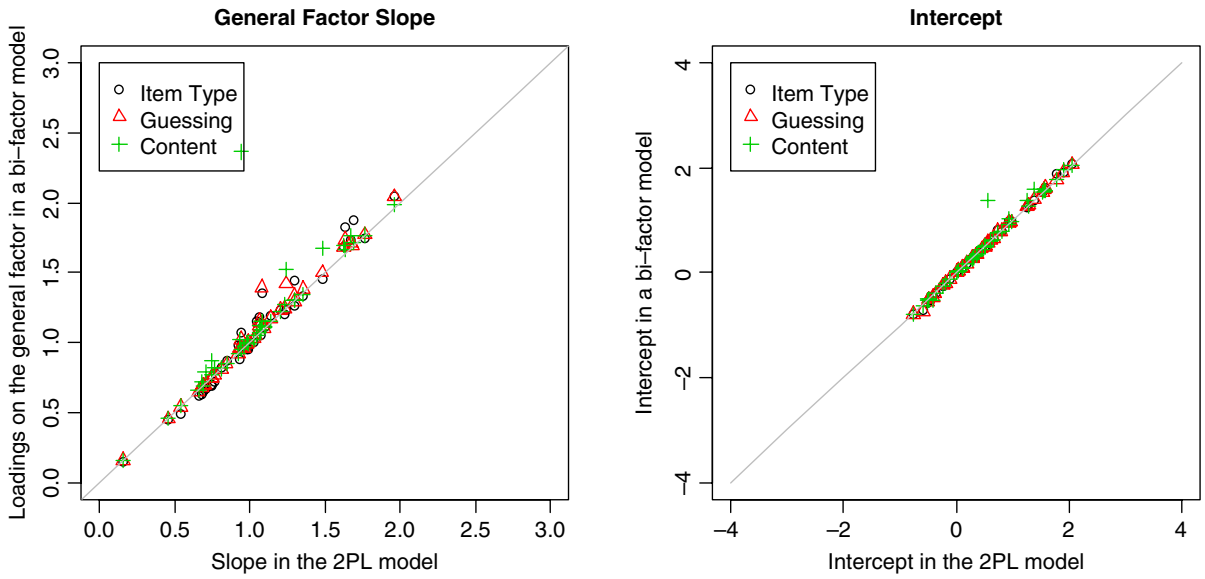

Figure 3 Comparisons of slopes and intercepts between the two-parameter logistic (2PL) and bifactor models for the verbal (VBL) jumpstart medium Form 1.

(Akaike, 1973; Schwarz, 1978) commonly used to identify the best fitting and most parsimonious model (e.g., Ip et al., 2013; Rijmen, 2010). With VBL, results indicated that item type appeared to be a meaningful secondary dimension; with QNT, while the AIC suggested some improvements over the 2PL, the BIC did not, casting doubt on the usefulness of any of the bifactor models investigated.

Tables 6 and 7 show the VBL and QNT loadings obtained with the three bifactor models for the medium Form 1 data. Item sets with more than one set member are shaded. When two sets are given next to each other, they are differentiated by different shading intensity. For example, Items 9 to 12 in Table 6 belong to the same set. Items 27 and 28 belong to one set, and Items 29 to 31 belong to another set.

For VBL, the loadings on the general factor were all reasonably high, with the exception of Item 10 (a very low discrimination item). Considering the item type model, the most promising of the bifactor models investigated, we observed that the secondary factor loadings were all smaller than the general loadings but still relatively high with RC items; the nonset-based items, TC and SE, had smaller loadings, with some positive as well as negative values. The consistency of the 2PL and item type bifactor parameters was very high, indicating that even though secondary factors were noticeable; taking them into account would not affect the overall VBL scores. Consistent with the AIC and BIC results, the other VBL bifactor models resulted in generally very low positive and negative loading, and a few items had very noticeable discrepancies between their 2PL and bifactor parameters. Therefore, we concluded that these models are not suitable. With QNT, the detailed results reinforced the interpretation of the AIC and BIC results. Therefore, we concluded that none of the bifactor models investigated are suitable. The same analyses were conducted on the VBL and QNT medium Form 2 data with similar results, leading and to the same interpretations.

Figure 3 illustrates the degree to which the VBL 2PL and bifactor general factor solutions are linearly related. While the item type bifactor solution fits the data better than the 2PL solution, the very close relationship found between the related parameters indicates that 2PL is nevertheless an appropriate model to use. Because the other VBL bifactor models do not fit the data better than the 2PL, further evaluation is irrelevant. For the same reason, we considered the QNT 2PL/bifactor plots irrelevant and did not provide them.

Steady State Confirmatory Factor Analyses

Given the results obtained with the jumpstart data, we focused our analyses on the most promising bifactor model identified: item type for VBL and content for QNT (although none of QNT bifactor models was satisfactory). Taking advantage of the ability of the full information factor analysis approach, we analyzed the full MST data. However, because of the

Table 8 Average Verbal (VBL) Bifactor Loadings Overall and by Multistage Testing (MST) Section

	Item type			
VBL	Gen	RC	TC	SE
MST	1.40	0.53	0.00	0.68
Routing	1.31	0.48	0.29	−0.09
Easy	1.54	0.56	−0.08	1.24
Medium	1.44	0.47	0.04	−0.01
Difficult	1.31	0.60	−0.24	1.41

Note. Gen = general factor; RC = reading comprehension; TC = text completion; SE = sentence equivalence.

Table 9 Average Quantitative (QNT) Bifactor Loadings Overall and by Multistage Test (MST) Section

	Content				
QNT	Gen	Alg	Ari	DA	Geo
MST	0.96	0.22	0.33	0.12	0.32
Routing	0.82	0.06	0.02	0.14	−0.07
Easy	0.99	0.22	0.24	0.08	0.36
Medium	0.82	−0.10	0.22	0.07	−0.06
Difficult	1.20	0.65	0.74	0.17	1.19

Note. Gen = general factor; Alg = algebra; Ari = arithmetic; DA = data analysis; Geo = geometry.

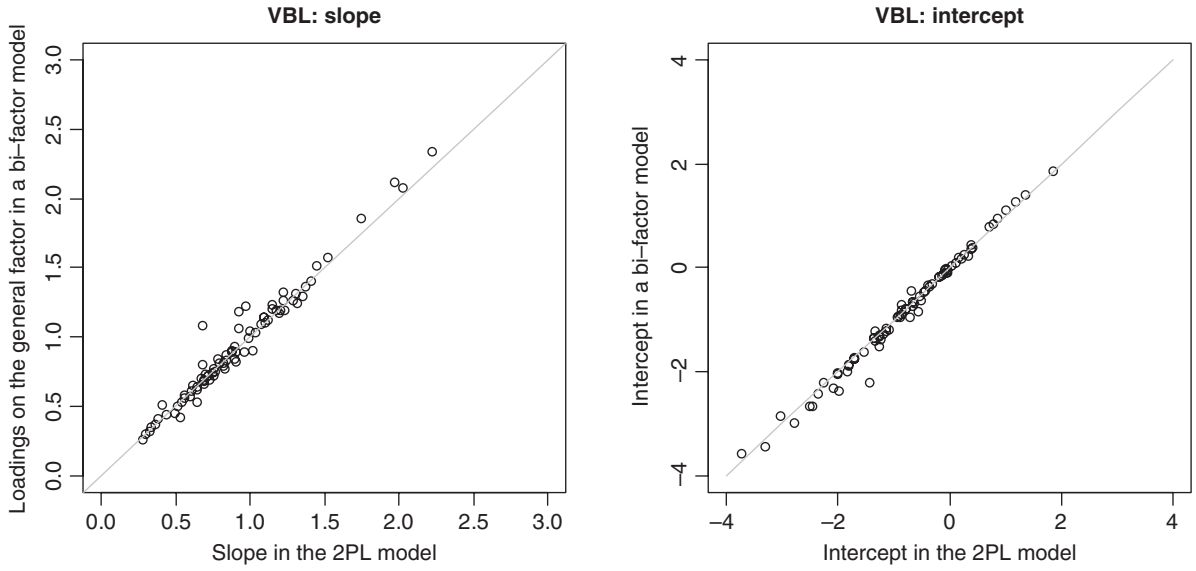

Figure 4 Comparison of the verbal (VBL) steady state unidimensional two-parameter logistic (2PL) and item type bifactor models' slopes and intercepts.

combination of relatively small sample sizes and missing data due to the MST design, these results should not be taken as definitive.

Tables 8 and 9 provide the average of the item loadings overall and by MST section. Tables 10 and 11 display the item loadings obtained. Figure 4 illustrates the degree to which the VBL 2PL slope and bifactor parameter intercepts are linearly related.

These results are very similar to the ones obtained with the jumpstart data. Also to be noted is the fact that the average loading values are very similar across MST sections. This was another result we hoped to see, as it provides evidence that the easy, medium, and difficult MST forms, despite their differences in difficulty, have the same factor structure that is appropriately modeled by the unidimensional 2PL.

Table 10 Item Type Bifactor Loadings for the Verbal (VBL) Steady Stage Multistage Test (MST) for Routing Easy, Medium, and Difficult 20-Item Sections

	Item type					Item type			
No.	Gen	RC	TC	SE	No.	Gen	RC	TC	SE
Item 1	1.31		−0.29		Item 41	1.28		0.41	
Item 2	2.07		1.21		Item 42	0.84		0.01	
Item 3	1.53		0.25		Item 43	0.75		−0.16	
Item 4	1.68		0.20		Item 44	1.70		−0.29	
Item 5	1.60		0.30		Item 45	1.00		0.48	
Item 6	1.67		0.06		Item 46	1.44		−0.20	
Item 7	0.78	0.56			Item 47	1.24	0.07		
Item 8	0.78	0.46			Item 48	0.79	0.40		
Item 9	0.57	0.06			Item 49	1.37	0.50		
Item 10	0.46	0.33			Item 50	0.90	0.25		
Item 11	1.33	0.63			Item 51	1.26	0.33		
Item 12	1.30	0.69			Item 52	1.42			−0.34
Item 13	1.95			0.13	Item 53	2.00			0.24
Item 14	1.69			−0.03	Item 54	11.23			15.00
Item 15	1.56			−0.16	Item 55	1.58			0.06
Item 16	1.51			−0.30	Item 56	1.77	0.63		
Item 17	0.73	0.54			Item 57	2.17	0.25		
Item 18	1.32	0.59			Item 58	1.64	0.34		
Item 19	0.81	0.44			Item 59	2.59	1.33		
Item 20	1.60	0.47			Item 60	1.68	0.63		
Item 21	2.46		−1.15		Item 61	1.18		−0.63	
Item 22	0.56		0.19		Item 62	1.16		0.64	
Item 23	2.15		0.27		Item 63	0.84		−1.08	
Item 24	1.76		0.11		Item 64	1.31		−0.40	
Item 25	2.02		−0.02		Item 65	2.19		−0.03	
Item 26	1.58		0.11		Item 66	1.02		0.08	
Item 27	0.93	0.25			Item 67	0.31	1.08		
Item 28	1.20	0.16			Item 68	1.58	0.56		
Item 29	0.93	0.58			Item 69	1.64	0.80		
Item 30	1.23	0.60			Item 70	1.13	0.25		
Item 31	−0.06	0.28			Item 71	1.11	1.22		
Item 32	2.91			0.52	Item 72	3.78			5.39
Item 33	3.48			0.38	Item 73	1.88			0.37
Item 34	1.72			3.90	Item 74	0.97			0.10
Item 35	0.81			0.16	Item 75	1.46			−0.24
Item 36	2.06	0.77			Item 76	0.82	0.62		
Item 37	1.99	1.40			Item 77	0.47	0.05		
Item 38	1.16	0.41			Item 78	1.92	0.08		
Item 39	0.90	0.79			Item 79	0.13	0.44		
Item 40	1.09	0.40			Item 80	1.31	0.87		

Note. Gen = general factor; RC = reading comprehension; TC = text completion; SE = sentence equivalence. Item sets with more than one set member are shaded. When two sets are given next to each other, they are differentiated by different shading intensity.

Summary and Conclusions

In light of the changes that were introduced by GRE VBL and QNT measures, it is important to verify the assumptions underlying the psychometric model used to implement the revision. As is true of any unidimensional IRT model, relying on the 2PL for scoring requires that (a) no significant level of local item dependence is present in any test forms, (b) a single ability dimension is sufficient to explain performance on the test, and (c) noticeable secondary dimensions, if they exist, are not sufficiently strong to bias the 2PL parameter estimates.

In this study, despite some limitations inherent to MST data collection, we were able to gather empirical evidence to support the assertion that the VBL reasoning and QNT reasoning GRE-MSTs satisfy the necessary requirements for using unidimensional 2PL IRT. More specifically, as a result of our analyses of domestic early MST medium forms and steady state MSTs, we found the following:

Table 11 Content Bifactor Loadings for the Quantitative (QNT) Steady Stage Multistage (MST) for Routing Easy, Medium, and Difficult 20-Item Sections

	Content						Content				
	Gen	Alg	Ari	DA	Geo		Gen	Alg	Ari	DA	Geo
Item 1	1		−0.1			Item 41	0.33		0.11		
Item 2	0.38			0.2		Item 42	0.36			0.17	
Item 3	1.74				0.1	Item 43	0.74				−0.33
Item 4	1.13	−0.36				Item 44	0.55	−0.48			
Item 5	0.7				0.05	Item 45	0.92		0.63		
Item 6	1.09			−0.1		Item 46	0.75	0.21			
Item 7	0.28	0.21				Item 47	1.19	−0.01			
Item 8	0.67		0.24			Item 48	0.62			−0.25	
Item 9	0.96		−0.04			Item 49	0.31		0.08		
Item 10	0.92			−0.18		Item 50	0.8		0.6		
Item 11	0.54			−0.1		Item 51	0.52			0.06	
Item 12	0.69				−0.54	Item 52	1		−0.42		
Item 13	0.71			0.18		Item 53	1.29	−0.54			
Item 14	0.82			0.15		Item 54	0.84			0.37	
Item 15	1.19			0.5		Item 55	0.86			0.35	
Item 16	0.56			0.43		Item 56	1.04			−0.29	
Item 17	0.83	0.26				Item 57	0.89				0.02
Item 18	0.76		−0.04			Item 58	1.19	0.34			
Item 19	0.77				0.1	Item 59	1.21		0.32		
Item 20	0.6	0.11				Item 60	1.07				0.12
Item 21	0.85	−0.07				Item 61	1.91		4.22		
Item 22	0.92	−0.05				Item 62	0.55		0.08		
Item 23	0.7	−0.02				Item 63	0.62	0.33			
Item 24	0.63				0.1	Item 64	1.33			−0.09	
Item 25	1.59		0.77			Item 65	1.37				0.54
Item 26	1.41				0.44	Item 66	1.32				0.09
Item 27	0.98			−0.44		Item 67	0.67		−0.05		
Item 28	1.56	0.97				Item 68	1.27			1.89	
Item 29	0.45		−0.09			Item 69	1.2	0.26			
Item 30	0.9			−0.2		Item 70	1.92	2.1			
Item 31	1.21		0.03			Item 71	0.89			−0.2	
Item 32	1.81			0.88		Item 72	1.17	0.05			
Item 33	0.78		0.76			Item 73	0.56				2.95
Item 34	1.2			0.48		Item 74	0.43			0.35	
Item 35	0.48			0.3		Item 75	1.12			0.45	
Item 36	0.76			−0.19		Item 76	0.99			−1.37	
Item 37	1.1	0.28				Item 77	1.92		0.26		
Item 38	0.74		−0.27			Item 78	2.16		0		
Item 39	1.01				0.53	Item 79	0.63		−0.1		
Item 40	0.7			−0.24		Item 80	1.9	0.51			

Note. Gen = general factor; Alg = algebra; Ari = arithmetic; DA = data analysis; Geo = geometry. Item sets with more than one set member are shaded.

- There was little to no evidence of local item dependence for either nonset-based or set-based items.
- There was evidence of a strong single factor for both VBL and QNT.
- The VBL bifactor model with secondary factors for each item type fitted the data somewhat better than the 2PL. However, the item parameters were effectively equivalent.
- The QNT 2PL model was more appropriate than the alternative bifactor models investigated.

In short, the results presented support the use of the 2PL IRT model as the psychometric model for GRE-MSTs.

However, as indicated throughout the paper, limitations inherent in the data collection for the early revised test have restricted the scope of this study. In the future, by purposefully using the variable sections, more extensive data may be collected and analyzed to confirm and expand on these early findings. In particular, while we have evidence that the IRT 2PL model fits the data sufficiently it is possible that its use both contributed to and obscured the multidimensional

results. The bifactor model, attractive for its simplicity, may also have been too constrained to capture some of the residual variance not accounted for by the model.

As growing numbers of international test takers have been taking the GRE-CAT and now the GRE-MST, the GRE program has incorporated item and test development guidelines and processes to ensure score comparability across subpopulations of test takers (Robin, 2014), and it is conducting research to identify potential sources of differential item functioning in order to further inform and guide practice (e.g., Oliveri, Lawless, & Robin, 2015). To contribute to this research and development effort and to confirm the validity of unidimensional scoring for international test takers, further assessment of test dimensionality is needed. Taking advantage of the flexibility of the test delivery design and the additional variable section, some of the limitations of the data collection used in this study can be overcome. Thus, with new data and new test forms, this study's finding may be expanded (a) to determine whether more complex unidimensional and multidimensional IRT models would provide more insight on the nature of the construct and (b) to evaluate the extent to which the dimensional structure of the test may vary across domestic test takers and test takers in relatively large-volume countries such as China or India.

Notes

1 The revised test also includes a nonadaptive analytical writing measure made up of two holistically scored essay prompts.
2 Specifically, as outlined by Kane (2006) the interpretive arguments for most scores include a scoring assumption that the raw responses are transformed appropriately to a score. The process for doing so with GRE-MST is to first obtain an IRT ability estimate based on the raw responses, which is then used to compute a number-right score on a reference form and finally converted to a score on the 130 to 170 reporting scale (Robin et al., 2014).
3 Before the launch of the revised test, field test data calibrations were conducted using both the 2PL and 3PL IRT models. The 2PL and 3PL models showed a similar level of model data fit, both overall and at the item level, as the routine inspection of the item residuals plots and statistics showed that equally few items needed to be removed from the calibrations. Since the launch of the revised test, through the calibrations of new items as well as the ongoing monitoring of postadministration data, the appropriateness of the 2PL model has been confirmed (Robin & Steffen, 2014).
4 The analogy and antonym items representing about half of the VBL test were replaced by text completion and sentence equivalence items; new multiple-choice, select-one-or-more-answers items, and numeric-entry items were introduced to the QNT test (Educational Testing Service, 2015a).
5 The GRE-PBT was restructured to include a new ANL measure to the already existing VBL and QNT measures.
6 While the test design and assembly blueprint were quite different, the GRE-CAT was designed to fulfill the same test specifications (including sampling of content and reliability) as the GRE-PBT. Although the new test was delivered on computer, there were very few modifications of the item formats used.
7 In 2012 the domestic test takers — U.S. citizens testing in U.S. test centers — represented 66.5% of the total population.
8 The jumpstart MSTs were assembled using pseudo item statistics obtained before the launch of the revised test, from various pretest data collections conducted along with the previous GRE-CAT tests. Therefore, the test takers' provisional scores could not be reported within the normal 2-week turnaround time. Simulation experiments showed that the quality of the pseudo IRT parameters would be sufficient to assemble MSTs with the desired measurement characteristics. After the jumpstart data were collected and the operational IRT parameters were determined, the quality of the jumpstart MSTs was confirmed. Since the jumpstart IRT calibrations were completed, the steady state MST assemblies have been conducted using precalibrated operational item parameters.
9 GRE test takers are first assigned two writing sections; they are then assigned two VBL, two QNT, and one unscored section in various ordering (Educational Testing Service, 2015c). All sections are timed: VBL sections are limited to 30 minutes, and QNT sections are limited to 35 minutes. The test delivery system can be configured to select the unscored section among a number of VBL, QNT or VBL and QNT sections. Unscored sections are assembled to similar specifications as the operational MST sections, so as to be undistinguishable from them, but they do not contribute to the reported scores. The primary purpose of the unscored section is to collect data for the pretesting of new items under operational testing conditions.
10 Each VBL and QNT MST form is constituted by two sections assigned: easy (R-E), medium (R-M), or difficult (R-D).
11 The program was developed by Krishna Tateneni, Gerhard Mels, Robert Cudeck, and Michael Browne. It is free software and can be obtained from http://faculty.psy.ohio-state.edu/browne/software.php.
12 VBL reading comprehension item sets and QNT data interpretation items sets use a common stimulus (ETS, 2015a).
13 We used the Matlab code BNLflirt (Rijmen & Jeon, 2013), which was later implemented in R as FLIRT, a free downloadable package from http://faculty.psy.ohio-state.edu/jeon/lab/flirt.php.

14 AIC = −2Loglikelihood + 2*number of parameters.
15 BIC = −2Loglikelihood + log(sample size)*number of parameters.
16 Factor loading matrices for higher factor models are available upon request.
17 AIC = −2Loglikelihood + 2*number of parameters.
18 BIC = −2Loglikelihood + log(sample size)*number of parameters.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrox and F. Caski (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261–280.

Briel, J. B., & Michel, R. (2014). Revisiting the GRE General test. In C. Wendler, B. Bridgeman, & C. Ezzo. (Eds.), *The research foundation for the GRE® revised general test: A compendium of studies.* Retrieved from http://www.ets.org/s/research/pdf/gre_compendium.pdf

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.

Dorans, N. J., & Kingston, N. M. (1985). The effect of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22*(4), 249–262.

Educational Testing Service. (2015a). *GRE revised General Test content and structure*. Retrieved from https://www.ets.org/gre/revised_general/about/content/

Educational Testing Service. (2015b). *GRE guide to the use of scores*. Retrieved from http://www.ets.org/s/gre/pdf/gre_guide.pdf

Educational Testing Service. (2015c). *Computer-delivered GRE revised General Test content and structure*. Retrieved from https://www.ets.org/gre/revised_general/about/content/computer/

Fox, J. (2010). *polycor: Polychoric and polyserial correlations.* R Package version 0.7-8. Retrieved from https://cran.r-project.org/web/packages/polycor/index.html

Gibbons, R. D., & Hedeker, R. D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423–436.

Gibbons, R. D., Immekus, J. C., & Bock, R. D. (2007). *The added value of multidimensional IRT models*. Center for Health Statistics, University of Chicago. Retrieved from http://www.healthstats.org/articles/NCI_Didactic_Workbook.pdf

Golub-Smith, M., & Moses, T. (2014). How the scales for the GRE® revised General Test were defined. In C. Wendler, B. Bridgeman, & C. Ezzo (Eds.), *The research foundation for the GRE® revised General Test: A compendium of studies* (3.3). Retrieved from http://www.ets.org/s/research/pdf/gre_compendium.pdf

Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research, 17,* 193–219.

Ip, E. H., Molenberghs, G., Chen, S. H., Goegebeur, Y., & De Boeck, P. (2013). Functionally unidimensional item response models for multivariate binary data. *Multivariate Behavioral Research*, *48*(4), 534–562.

Jeon, M., & Rijmen, F. (2010, April). *Assessing differential item functioning for testlest-based tests using the bifactor model*. Paper presented at the annual meeting of the National Council of Measurement in Education, Denver, CO.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2014). Flexible item response theory modeling with FLIRT. *Applied Psychological Methods, 38,* 404–405.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*(2), 147–154.

Liu, O. L., Bridgeman, B., Gu, L., Xu, J., & Kong, N. (2015, March). Investigation of response changes in the GRE revised General Test. *Educational and Psychological Measurement.* Advance online publication. doi:10.1177/0013164415573988. Retrieved from http://epm.sagepub.com/content/early/2015/02/26/0013164415573988.abstract

Oliveri, M. E., Lawless, R., & Robin, F. (2015). *An exploratory analysis of differential item functioning and its sources across multiple GRE test taker populations*. Unpublished report. Princeton, NJ: Educational Testing Service.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*(3), 200–219.

Powers, D. E., & Swinton, S. S. (1981). Extending the measurement of graduate admission abilities beyond the verbal and quantitative domains. *Applied Psychological Measurement, 5*(2), 141–158.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696.

Rijmen, F. (2009). *Efficient full information maximum likelihood estimation for multidimensional IRT models* (Research Report No. RR-09-03). Princeton, NJ: Educational Testing Service. 10.1002/j.2333-8504.2009.tb02160.x

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361 – 372.

Rijmen, F., & Jeon, M. (2013). *BNLflirt: Flexible item response theory modeling with BNL*. Matlab file exchange.

Robin, F. (2014). Fairness and group performance on the GRE® revised General Test. In C. Wendler, B. Bridgeman, & C. Ezzo (Eds.), *The research foundation for the GRE® revised General Test: A compendium of studies* (6.6). Retrieved from http://www.ets.org/s/research/pdf/gre_compendium.pdf

Robin, F., & Steffen, M. (2014). Test design for the GRE® revised General Test. In C. Wendler, B. Bridgeman, & C. Ezzo (Eds.), *The research foundation for the GRE® revised General Test: A compendium of studies* (3.3). Retrieved from http://www.ets.org/s/research/pdf/gre_compendium.pdf

Robin, F., Steffen, M., & Liang, L. (2014). The multistage test implementation of the GRE revised General Test. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 325 – 341). Boca Raton, FL: CRC Press.

Stricker, L. J., & Rock, D. A. (1985). *Factor structure of the GRE General Test for older examinees: Implications for construct validity* (GRE Board Research Report No. 83-10R; ETS Research Report No. RR-85-09). Princeton, NJ: Educational Testing Service. 10.1002/j.2330-8516.1985.tb00094.x

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461 – 464.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237 – 247.

Thissen, D., & Wainer, H. (2001). *y*. Mahwah, NJ: Lawrence Erlbaum.

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34,* 421 – 459.

Yen, W. M. (1984a). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement*, *21,* 93 – 111.

Yen, W. M. (1984b). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125 – 145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187 – 213.

Suggested citation:

Robin, R., Bejar, I., Liang, L., & Rijmen, F. (2016). *Dimensionality analyses of the GRE® revised General Test verbal and quantitative measures* (GRE Board Research Report No. 16-02, ETS Research Report No. RR-16-20). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12106

Action Editor: Brent Bridgeman

Reviewers: This report was reviewed by the GRE Technical Advisory Committee and the Research Committee and Diversity, Equity and Inclusion Committee of the GRE Board.