JOURNAL OF APPLIED MEASUREMENT, *10*(2), 170-184

Constructing One Scale to Describe Two Statewide Exams

Insu Paek
*Harcourt Assessment**

Deborah G. Peres

Mark Wilson
University of California at Berkeley

This study applies two approaches in creating a single scale from two separate statewide exams (Golden State Math Exam and California Standard Math Test) and compares some aspects of the two statewide tests. The first analysis involves a sequence of unidimensional Rasch scalings, using anchored items to scale the two tests together. The second analysis employs a 2-dimensional Rasch scaling using previous unidimensional analysis results to link the scales. The linking facilitates the investigation of their measurement properties of the two exams and is a basis for combining items from both exams to develop a more efficient testing program. The results of the comparisons of the two statewide exams based on the linking are shown and discussed.

* Now at Educational Testing Service

Requests for reprints should be sent to Insu Paek, Educational Testing Service, Rosedale Road, MS-O2P, Princeton, NJ 08541, USA, e-mail: ipaek@ets.org.

Two large scale testing programs in California, the Golden State Examinations (GSE) and the California Standards Test (CST) have some common subjects (e.g., Algebra, Biology). The GSE is an honor-awarding test. It is not mandatory. It classifies students into six levels of performance. The top three levels (6, 5, 4) are entitled "high honors", "honors", and "school recognition". Students who score at these levels on a number of exams are eligible for honors diplomas; scores at levels 3, 2 and 1 do not qualify for honors. Therefore, in GSE, the cutoff between the performance levels 4 and 3 is very important.

The CST is part of the Standardized Testing and Reporting system and is mandatory for all students in the state, including those who have just taken, or are about to take, the GSE. The CST is used as part of schools' Academic Performance Index (API) on which schools and teachers are judged, and scores are also reported at the individual student level. The CST is a high-stakes test compared to the GSE.

Many students take both tests within a few months of each other. This raises the question of whether there is overlap between the two tests; the presence of overlap could mean there is an unnecessary testing burden on students. Might it be possible to maintain the quality of testing (i.e., identify honors students and gather information about all students' academic performance) with only one test, or with two tests whose combined times is less than the current combination? For example, since all students take the CST, might it be possible to use information from the CST to supplement a shorter GSE. The reduction of testing time question can also be rephrased as "How much overlap between the two tests?" or "Are all items necessary in both tests?" One way to investigate these questions is to examine the measurement properties of the two tests, (i.e., comparison of item difficulties and information across the two tests). To this end, construction of one scale for the two tests is necessary. This study illustrates how we achieved the construction of one scale, using unidimensional item response theory (IRT) and multidimensional IRT approaches. We also show and discuss some aspects of the two statewide exams based on the calibrations and linking.

Data Set

We chose one of several subjects common to the GSE and CST: high school math. The GSE test is called the High School Mathematics test (GSE math) and the CST test in this subject is the 11th Grade Mathematics test (CST math). Both are cumulative examinations covering three years of mathematics: Algebra, Geometry, and Algebra II. The GSE math has 40 multiple-choice (MC) items and 2 open-ended (OE) questions that are each scored with a four-point rubric. The CST math has 65 multiple-choice items only. The data editing led to 8053 students who took the both tests.

In the GSE math, there were a priori item scoring weights for MC and OE used on the operational test, based on decisions by the California Department of Education. We used the same relative weights for our analyses. The scoring weights for MC were 0 and 1 for the incorrect and correct answer (as usual); the scoring weights for OE were 0, 3, 6, and 9 for the lowest to the highest categories on the rubric. This is equivalent to the weighting scheme of 0.69 : 0.31 for MC versus OE in terms of the raw total test scores on the GSE math. (If the OE had been scored 0, 1, 2, and 3 as usual, the resulting weighting scheme would be 0.87: 0.13 for MC versus OE in terms of the raw total scores. So more weight was given to the OE items than in the typical OE scoring scheme of 0, 1, 2, 3). All items for the CST math were scored 0 and 1 for the incorrect and the correct answer.

Method

Item Response Model

It is well known that IRT modeling is useful for dealing with complex measurement problems. It can relatively easily handle linking problems in testing (Hamblton, Swaminathan and Rogers, 1991). For the construction of one scale in our analyses, we chose an IRT model called the multidimensional random coefficient multinomial logit model (Adams, Wilson, and Wu, 1997). The

multidimensional random coefficient multinomial logit model (MRCMLM) is an extension of random coefficient multinomial logit model (Adams and Wilson, 1996), a Rasch type generalized item response model. It can fit many existing IRT models, for example, the simple logistic model (Rasch, 1960), the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982), the ordered partition model (Wilson, 1992), the linear logistic test model (Fisher, 1983), the multifacet model (Linacre, 1994), and the multidimensional versions of all these models.

The MRCMLM also allows user-specified IRT modeling through importing a matrix that is called the design matrix. Another relevant feature for our analysis is its capability to accommodate a priori given item weights (not empirically estimated) in the design matrix; this is done through its scoring function. In this sense, MRCMLM is a useful tool when items in a test are given a priori weights by content experts, e.g., from a weight setting committee.

The MRCMLM was substantiated by the program, ConQuest (Wu, Adams, and Wilson, 1998). All the estimations of the item response models for the unidimensional and the multidimensional approaches in this paper were done using ConQuest. The general form of the item response models used for our analysis here is:

$$P(X_{nik}=1|\Theta_n,\xi) = \frac{\exp[\mathbf{b}'_{ik}\Theta_n+\mathbf{a}'_{ik}\xi]}{\sum_{k=1}^{K_i}\exp[\mathbf{b}'_{ik}\Theta_n+\mathbf{a}'_{ik}\xi]}, \qquad (1)$$

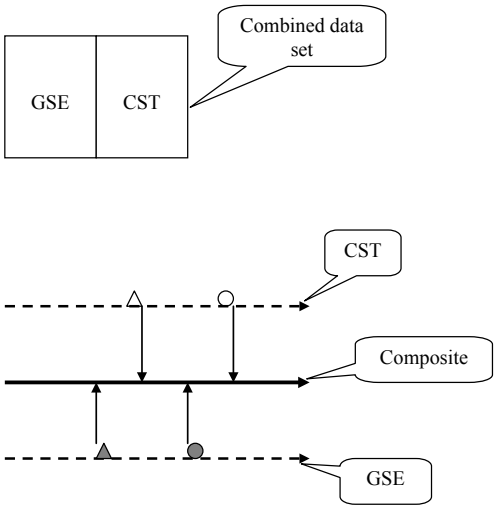
where $X_{nik} = 1$ if person n's response to item i is in category k, or 0 otherwise ($1 \le i \le I$, $1 \le k \le K_i$, $1 \le n \le N$, and X_{ni1} is fixed to zero as a reference category for model identification); Θ_n is a d x 1 proficiency (or ability) parameter vector of a person n ($1 \le d$ (dimension) $\le$ D); $\mathbf{b}'_{ik}$ is a 1 x d scoring vector for category k of item i ; ξ is a p x 1 item parameter vector; and $\mathbf{a}'_{ik}$ is a 1 x p vector to specify linear combination of p elements of ξ for each response category.

Note that Θ_n is a random parameter vector and ξ is a fixed unknown parameter vector. The elements of Θ_n are assumed to be independently and identically distributed from a multivariate normal (MVN) distribution. The item parameter of ξ is estimated by marginal maximum likelihood (MML) method. For Θ_n estimation, maximum likelihood estimate, expected a posterior (EAP), and Warm's (1989) weighted likelihood estimate are available in ConQuest.

Composite Unidimensional Calibration

There are many IRT-based methods available for test linking and achieving a common scale, for example, test response surface minimization method, mean-sigma method, fixed anchor item method, and concurrent calibration (see Kolen and Brennan, 2004; von Davier and von Davier, 2004 for detailed descriptions). Our situation is a set of common matching cases (examinees) but two different tests. Given the common cases between the two tests and the availability of the two test response data sets, one of the obvious ways to achieve one scale is concurrent calibration; in other words, combine the two test data sets as one and calibrate it as if all responses were part of one big test. Since the test blueprints are different for the two tests, the scale we are constructing in this way can be called a composite scale. This composite scale can be interpreted as representing the intersection of the curriculum objectives that are shared by both the GSE math and the CST math. It provides a ground for comparison of the two tests as such.

Conceptual representation of this linking approach is shown in Figure 1. The rectangular boxes adjacent to each other without any space represent the data matrix structure for unidimensional composite calibration. The vector-like figure is a conceptual diagram. Dotted lines represent the scales of the GSE math and the CST math. In this composite unidimensional approach, their scales are transformed onto the composite scale such that all item and case estimates from both tests are aligned on the composite scale.



Note: white triangle and circle mean an item and a case estimate in CST and gray triangle and circle means an item and a case estimate in GSE.

Figure 1. Conceptual representation of composite undimensional scaling.

Sequential Unidimensional Calibration

To express the scale in terms of either of the original test metrics, one can use the unidimensional calibration procedure sequentially. In our analyses, either the GSE math and the CST math scale could be chosen as the basis for the common scale; we chose the GSE math scale because educators and administrators in the California Department of Education were more familiar with the GSE scale and because work had been done to provide qualitative descriptions (e.g., "honors" level performance and what skills that performance requires) for different levels on the GSE math scale.

Our construction of a unidimensional GSE-based scale starts with the calibration of the GSE math test alone. This first step is to obtain the item and the case estimates for GSE math. The second step involves the use of the combined data from GSE math and CST math. In this second step, the unidimensional calibration with the combined data set is performed while the GSE items are anchored to the values estimated in the first calibration. The second step is in order to obtain the CST item estimates. The item estimates of the CST math test produced in the second step are on the GSE-based scale, so they can be interpreted and compared with the GSE item estimates. However, a third step is necessary to obtain CST math case estimates on the GSE math scale. The case estimates produced in the second step are not the ones that are desired because they are based on responses to both tests, not only on CST math responses; what is wanted is the case estimates based only on the CST math responses. In this last step (the third step), only the CST test data set is used, and case estimation is performed while the CST items estimates are anchored using estimates from the second step. The procedure of this sequential unidimensional calibration is summarized in Figure 2. Conceptual graphic representation of the procedure is shown in Figure 3.

We can think of the first step as creating a line for the GSE math and all the item and case estimates are aligned on the line. In the second step, we create a composite such that the locations of the GSE items on the composite line are the same as those on the GSE line and the CST items are located onto this composite line. In the third step, we create a line for CST such that the locations of the CST items on the CST line are the same as those on the composite line and all case estimates are located on this new CST line.

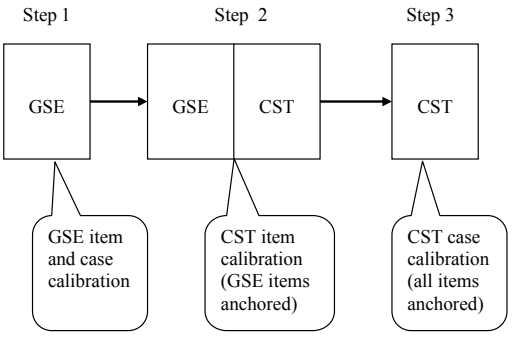


Figure 2. Procedure for the sequential undimensional (1-D) scaling: Construction of 1-D GSE based scale.

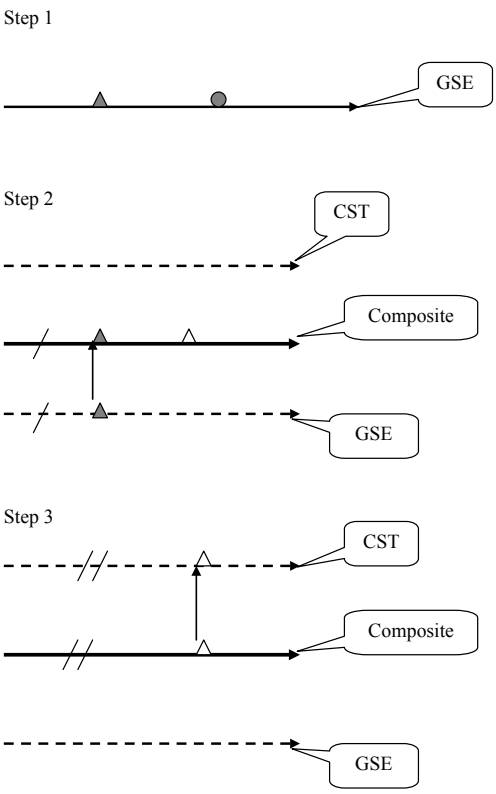


Figure 3. Conceptual representation of the undimensional sequential scaling.

Multidimensional Calibration

Although the GSE math and the CST math share common content matter and are meant to cover the same three years of math as taught in California, the test blueprints are not the same and the test items are different from each other. Therefore, precisely speaking, the tests can be viewed as measuring two separate, but strongly related, constructs. Under this multidimensional perspective on the two tests, one can apply multidimensional IRT model.

One of the ways to express the 2-D calibration results onto the GSE-based 1-D calibration scale is to anchor the 2-D population model parameters using the population parameter estimates from the GSE-based 1-D calibration. The anchored 2-D population structure is:

$$\Theta \sim MN\left[\begin{pmatrix}\mu_G * \\ \mu_C *\end{pmatrix},\begin{pmatrix}\sigma_G^2 & \text{cov} \\ \text{cov} & \sigma_C^2 *\end{pmatrix}\right], \qquad (2)$$

where Θ is a 2 x 1 latent ability vector, "~" means "follows", *MN* represents multivariate normal distribution, and "*" means the parameter is fixed. In our analyses, this anchored 2-D population structure was used to link the GSE and CST tests in the 2-D calibration and to link between the 2-D and the GSE-based (sequential) 1-D analysis. The μ_G * is the model constraint in the GSE part in the 2-D analysis and it is fixed to zero. The same population model constraint was used for the 1-D GSE calibration. μ_C * and σ_C * are the population model estimates of the CST test from the GSE-based 1-D calibration.

The sequential unidimensional calibration followed by this 2-D calibration using the anchored population structure above makes a common scale across the two tests and across the different dimensional approaches in the sense that all the constructed scales share the same origin and the unit of measurement. The comparisons of these calibration results are shown in the next section.

Information Functions

Following the methods of Samejima (1969) and Bock (1972), test information, item information, and item category information functions were used to compare the GSE math and the CST math.

For item information,

$$I_i(\theta) = \sum_{k=1}^{K}\left[\left(\frac{\partial P_{ik}(\theta)}{\partial\theta}\right)^2 \Big/ P_{ik}(\theta) - \frac{\partial^2 P_{ik}(\theta)}{\partial\theta^2}\right], \quad (3)$$

where $P_{ik}(\theta)$ is $P(X_{nik} = 1|\Theta_n,\xi)$ with Θ_n a scalar of θ now. For test information,

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta). \qquad (4)$$

For item category information,

$$I_{ik}(\theta) = I_i(\theta)P_{ik}(\theta). \qquad (5)$$

The above functions can still be used to summarize the marginal information for each dimension (each test) in our multidimensional analysis where an item loads onto only a single dimension.

Results

Sequential Unidimensional Calibration

To facilitate the comparisons of item estimates and case distributions of the CST math and the GSE math, Figure 4 shows an item map that includes all item locations (difficulty estimates) from both tests on the common logit scale created using the sequential GSE-based 1-D calibrations. The map also shows the case distributions. The GSE performance cutoff scores are shown to facilitate the comparability. The case distributions plotted in the item map are based upon plausible values. A plausible value is a random draw from a person's posterior distribution. Plausible values rather than best estimates such as EAPs or MLEs are better choice to represent a population (Mislevy, 1984). Plausible values can be of use to better estimate population characteristics as a secondary analysis data (see, e.g., Adams, Wilson, and Wu, 1997).

On the item map, threshold estimates are used to represent the polytomous nature of the GSE OE items. The threshold is defined as the point on the logit latent scale that corresponds to probability of 0.5 in the cumulative category probability function (Masters, 1988). It is another way to express polytomous item parameters and it has the property that it should be always sequentially ordered following the ordinal level of item category.

In Figure 4, the GSE OE questions appear to work well to differentiate the higher performance levels (4, 5, 6). Their lowest thresholds are located in the upper half of level 3 which is close to the most important cutoff in the GSE math (the cut between 3 and 4); the higher OE thresholds are in the honors-eligible levels of performance (4, 5, and 6). On the other hand, about 68% of MC item (27 items) in GSE are located below zero logits. They range from the middle of level 3 to the lowest level. Given that the most important cutoff for GSE is between levels 3 and 4, it seems that those MC items below about zero logits are not optimal for the GSE purpose of identifying honors students.[1] Also these items are redundant in the sense that many CST items are found in the same ability (logit) range.

Overall, the MC items in CST in Figure 4, compared to GSE MC items, cover a wider range of ability. This is what one would expect since the CST is mandatory for all students and must measure them all well, while the GSE is supposed to differentiate honors students from those that are not honors eligible (and differentiate levels of honors). Despite the wider range of the CST math, the locations of the CST items in Figure 4 show that the CST math differentiates students in the middle and lower ability range (i.e., GSE levels 3, 2 and 1) better than in the higher ability range; more items are located in the middle and lower ability range than in the high ability range. The CST math is meant to provide information at the individual level as well as the school level, for students at all levels. The CST math could do a better job of measuring the higher performing students by including more of the difficult items, i.e., those items located above 0.0 or 0.3 logits.

The test information curves for GSE and CST are shown in Figure 5. The solid lines are the test information curves and the dotted vertical lines are the GSE cutoffs, which divide the theta scale into the six GSE performance levels.

The CST information is spread wide and flat compared to the GSE information; the latter is highly peaked between zero and one in the logit scale.

1 One reviewer pointed out that there were few GSE items in the higher difficulty range, also asking if there were enough data for calibration of the lower difficulty items. The proportion (or percentage) of the correct answer (i.e., classical test theory item p-value) can be used as a quick check for diagnosing a potential item calibration problem. Very or extremely high or low p-value items are likely to be calibrated with relatively less precision. The p-values for the highest and the lowest difficulty GSE items (item number 17 and 23 respectively) were 20.29 and 91.31. The GSE item number 23 was a very easy item. It showed about 1.6 times larger approximate standard error estimate (0.04) than the other GSE items. The p-values for the highest and the lowest difficulty CST items (item number 29 and 53 respectively) were 30.90 and 98.36. The CST item number 53 was an extremely easy item. Compared to the other CST items, its approximate standard error estimate was 0.09, which was 4.5 times as large as the smallest approximate standard error estimate and 2.5 times as large as the second largest approximate standard error estimate.

GSE PLVL	GSE MC	GSE OE1	GSE OE2	CST MC	Logit	GSE Histogram	CST Histogram
					3.5		
							<
					3		
							<
							<
					2.5		<
							<
							<
							<
					2	<	<
							<
							<
							<
						<	X
					1.5	<	X
	17					<	XX
						<	XX
						X	XX
						X	XX
6	19		42.3		1	X	XXXX
				29		XXX	XXXX
		41.3		35		XXX	XXXX
5	24			16 57		XXXXX	XXXXX
		41.2	42.2			XXXXXXX	XXXXX
	20			49	0.5	XXXXXXXX	XXXXXX
4	39			64		XXXXXXXX	XXXXX
	16 40			22 58		XXXXXXXXXX	XXXXXXX
	4		42.1	7 56		XXXXXXXX	XXXXXXXX
	5 12 14			11 14		XXXXXXXXXX	XXXXXXX
	13 28 37	41.1		21	0	XXXXXXXX	XXXXXX
	11 18			15 18 37 41 43 44 54		XXXXXXXXX	XXXXXXXX
	10			5 13 62		XXXXXXXXX	XXXXXX
	30 32			9		XXXXXXXXXXX	XXXXXXXX
	2 8 15 31			32		XXXXXXX	XXXXXX
3	1 9 27 33 35 36			24 39 50 60	-0.5	XXXXXX	XXXXXXX
	21 29			63 65		XXXXX	XXXXXXX
	3 6 7			40 61		XXXXXX	XXXXX
				4 33 38 47		XXXX	XXXXX
	26 34			19 28 42		XXX	XXXX
2				10	-1	XX	XX
	38			17 26 59		X	XX
	25					<	XX
	22			2 31 45		X	X
				25 51		<	X
				52	-1.5	<	X
				6 30 46		<	<
				27		<	<
				1 3			<
				12 23 48			<
					-2	<	<
				8 34			<
				20			<
				36 55			<
							<
	23				-2.5		
							<
							<
					-3		<
					-3.5		XX
					-4		
				53			
1					-4.5		

Note: "x" and "<" in GSE and CST histograms represent 50 and under 50 respectively.

Figure 4. Sequential unidimensional calibration.

The highest information point for CST is near the GSE cutoff score between the performance level 2 and 3 (the second dotted line). For GSE the highest point is near the GSE cutoff score between the performance level 4 and 5 (the fourth dotted line). The most important GSE cutoff line (the third dotted line) to differentiate honors students and non-honors students has the second highest information point among the five GSE cutoff lines.

Previously it was mentioned that the two GSE OE items appear to work well. To better investigate their item properties, their item category characteristic curves, item information, and item category information plots are shown in Figures 6 and 7. Again, the dotted vertical lines are the GSE performance level cutoff lines.

Category 3 in GSE OE item 41 shows low probability compared to the other categories in Figure 6 (a). This suggests that the item could function just as well with only three categories as it does with four, if category 3 were collapsed into an adjacent category. A student having ability at about the GSE cutoff between the level 3 (non-honors) and level 4 (honors)—the third dotted line—has the highest likelihood of scoring category 2 (i.e., score of 3 on (0, 3, 6, 9) scoring scheme) in the two OE items. The item information plots in Figure 7 for the two OE items look very similar to each other. The highest peaks for the two OE items are located near the cutoff between the performance level 4 and 5 (the fourth dotted line). The third GSE cutoff line (between non-honors and honors) has the second highest information among the five cutoff lines. These are consistent with the observations about the GSE test information from Figure 5.

If a more efficient testing program is desired in terms of the reduction of the testing time, one way is to combine items from the GSE and CST tests. It was observed from Figures 4 and 5 that the GSE test is more effective for higher ability students and the CST works better for middle and lower ability students. Based on this observation, non-optimal GSE MC items that are located in same ability range as many CST MC items could be taken out and the rest of GSE MC items (e.g., GSE MC items above about zero logit and the 2 OE items) could be combined with the CST MC

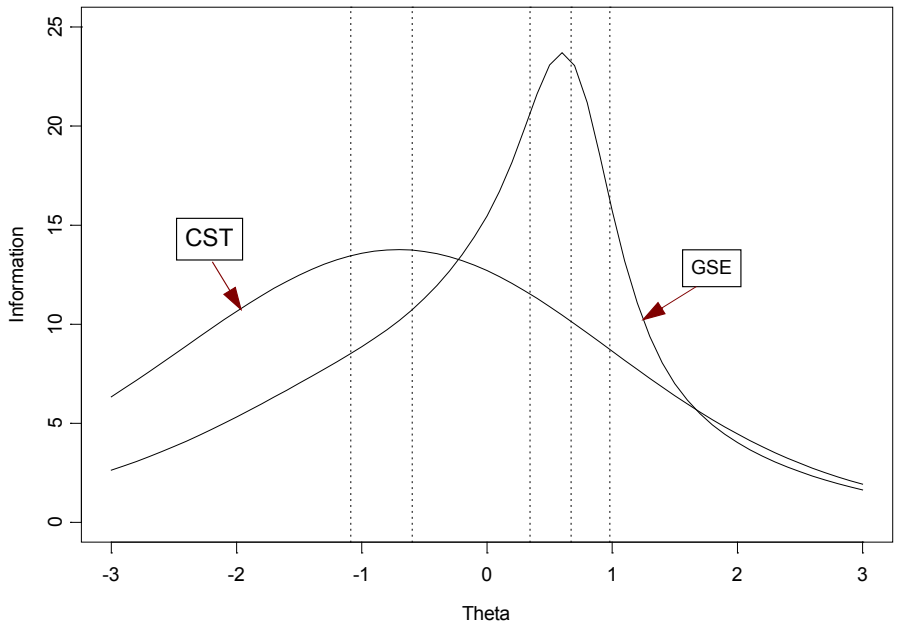


Figure 5. Test information plots for GSE and CST.

items below about zero logit. The number of the GSE MC items at or above zero logit is 13 and the number of the CST MC items below zero logit is 52. When those selected MC items from both tests are combined with the two GSE OE items, the total number of items in the combined test is 67. This amounts to about 37.4% of the test length reduction compared to the current two separate test administration. The test information of this artificial combined test is shown in Figure 8.

The combined test information is higher than the current two exams in the region between –1 and 2 logit. It achieves the shortened test length and better information.

Based upon the well-functioning behavior of the two GSE OE items at the middle and higher ability range, adding only the two OE items in GSE to the CST test could be an option as well. The test information of this combination test turned out to be virtually the same as the previous combined test information; overlaying this information does not show any difference from the previous combined test information shown in Figure 8, so it was not shown. The absolute mean difference between the previous combination and this combination test information was 0.023 for proficiency ranging from –3 to 3 by 0.1 increase. The maximum and the minimum absolute differences were 0.059 and 0.00007 respectively. This combined test has 65 CST item + 2 GSE items. The previous combined test had 13 GSE MC items + 52 CST MC items + 2 GSE OE items. Both combined tests show the same amount of test length reduction: 37.4% compared to the current two separate test administration. And either way of combining items showed more information and shortened test length, thereby the reduction of testing time for the students who take both tests.

Multidimensional (2-dimensional) Calibration

The results of 2-D calibration are shown in Figure 9. They are also summarized in an item map in the same manner as in Figure 4.

The item map in Figure 9 shows very similar trends to the sequential unidimensional calibration shown in Figure 4. In terms of item estimates, there were only slight differences. The mean absolute differences for GSE and CST item estimates between the sequential 1-D analysis and the 2-D analysis were 0.008 and 0.047 respectively. The minimum and maximum absolute differences for GSE were 0.001 and 0.015; for CST, 0.0006 and 0.270. The sequential unidimensional approach and the 2-D approach turned out almost the same in terms of item difficulty locations. Case histogram representations, and information analysis were also very similar to the sequential 1-D analysis. Therefore, the comments and interpretations made for Figures 4 through 8 and suggestions about the reduction of testing time in the unidimensional approach can be applied to this 2-D calibration results without much loss of generality, so they are not repeated in here.

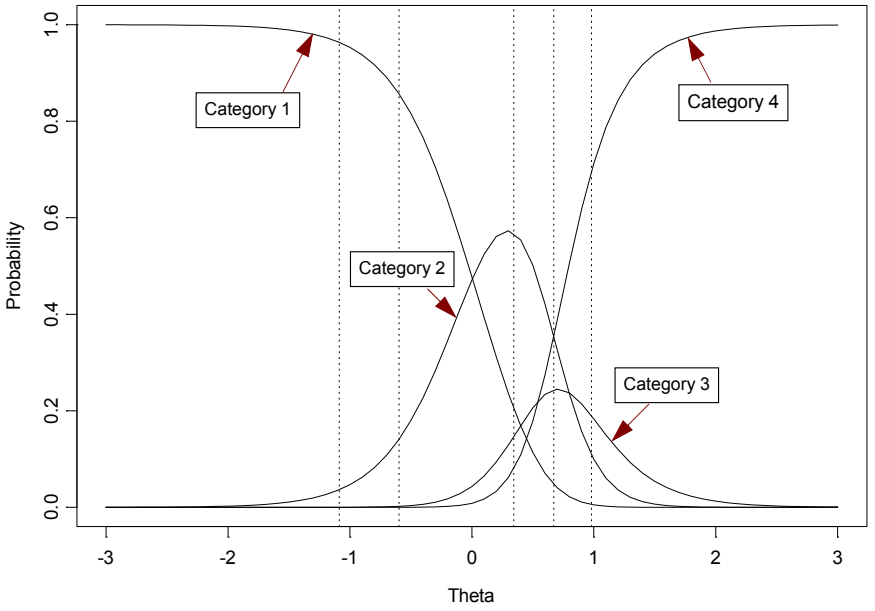
Case Summary Statistics and Model-to-Data Fit

Summary statistics for case representation in the unidimensional and multidimensional approaches are shown in Table 1.

As was noticed in Figures 4 and 9, the CST math shows a case distribution that is about the twice as large as on the GSE math in its standard deviation. The direct dimension correlation between GSE and CST from the 2-D calibration turned out 0.743. This direct latent dimension correlation has the good property that it is free of the noise from the measurement error of individual case estimation, and in general this direct correlation is larger estimate than the one calculated using the individual cases' best estimates. In this sense, it is a disattenuated latent correlation (Wang, 1999).

Table 2 shows the model fit for the 1-D analysis and the 2-D analysis, using the deviance—the value of the objective function at its minimum achieved under the default convergence criteria in the program ConQuest. The deviances in the first two columns are from the GSE and the CST calibrations respectively. The third column is the sum of the deviances in the first two columns. Note that the number of parameters in the 2-D analysis is smaller than that of the 1-D sequential separate calibration, due to the partially fixed

(a) GSE OE item 41



(b) GSE OE item 42

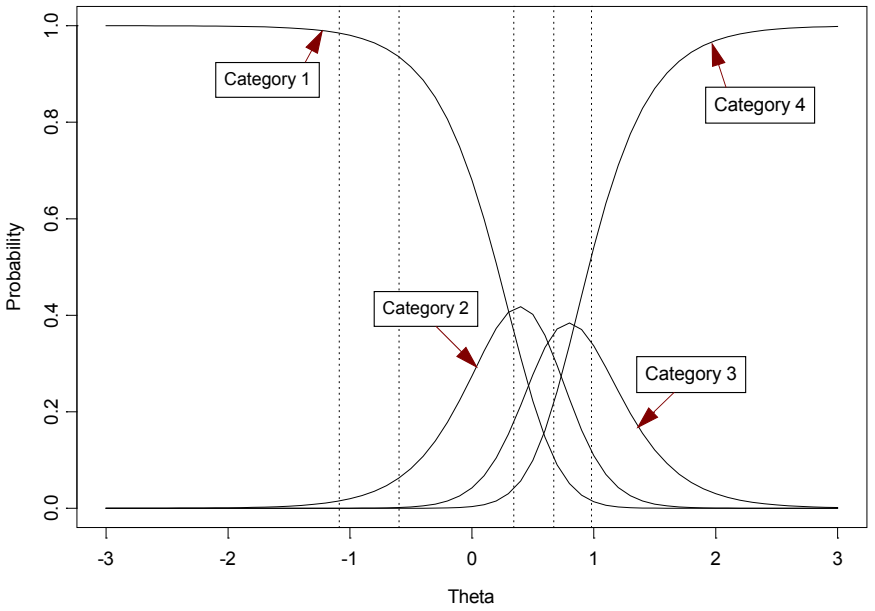
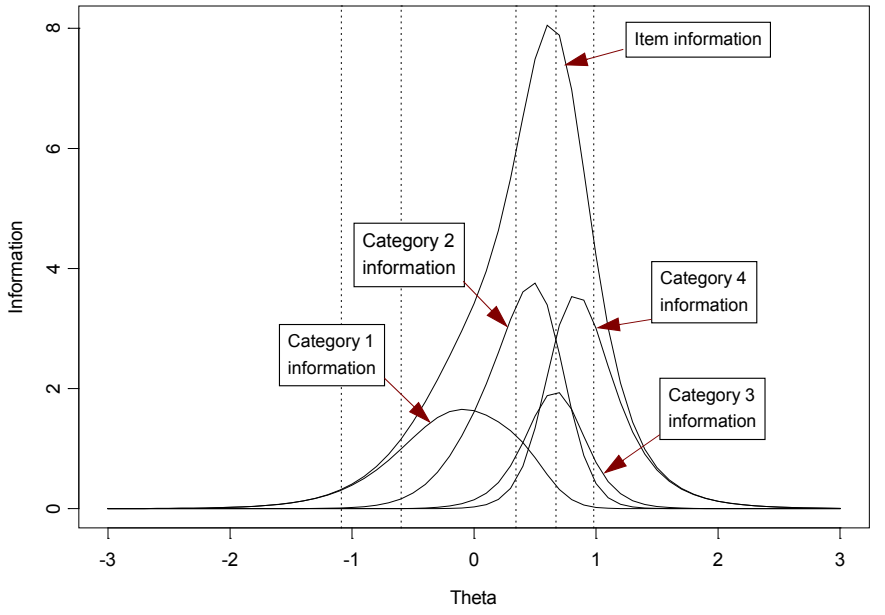


Figure 6. Item category characteristic curves for two GSE OE ítems.

(a) GSE OE item 41



(b) GSE OE item 42

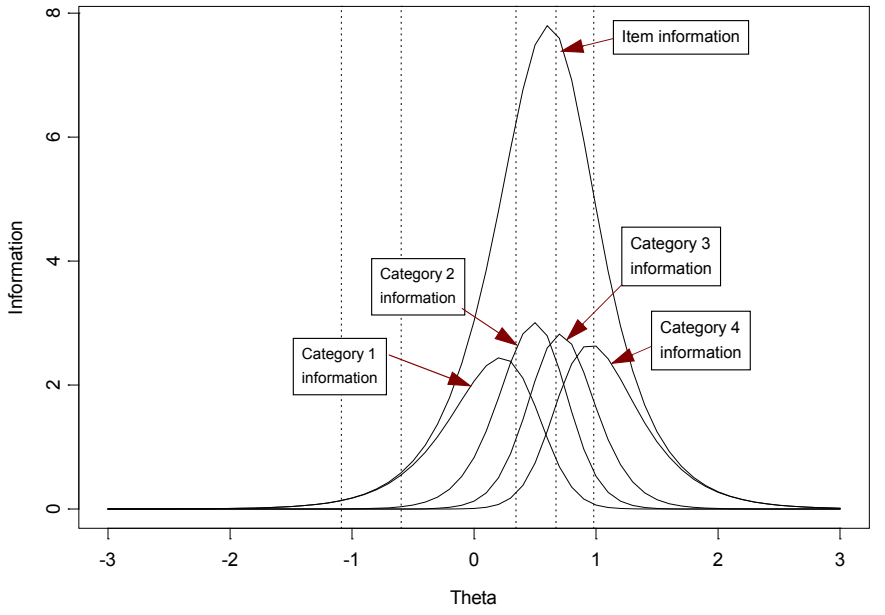


Figure 7. Item information and item category information for two GSE OE items.

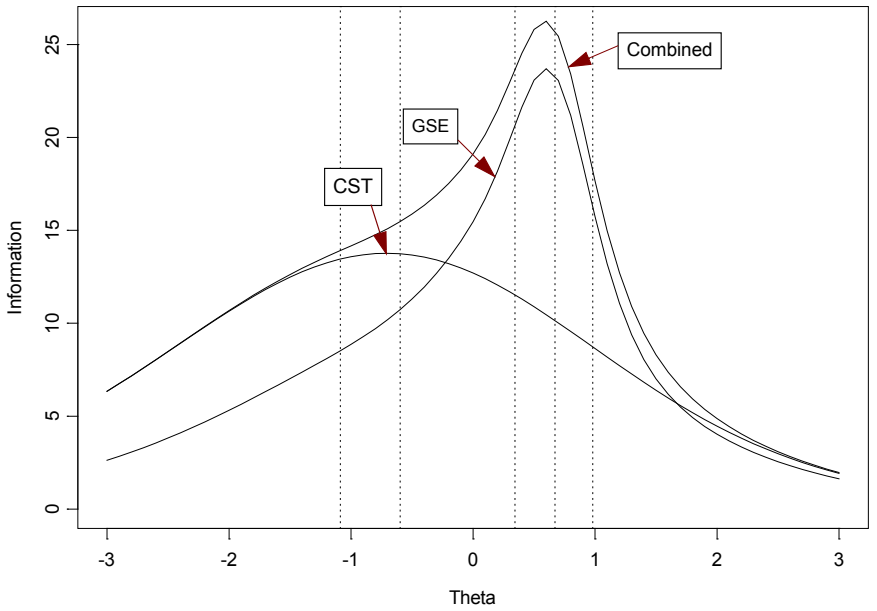


Figure 8. Test information for the combined test.

Table 1

Case Summary Statistics

	Sequential 1-D		2-D	
	GSE	CST	GSE	CST
Mean	0*	–0.025	0*	–0.025*
SD	0.556	0.956	0.559	0.956*

Note: Disattenuated correlation between GSE and CST from the 2-D calibration was 0.743. "*" represents an anchored parameter value.

Table 2

Model Fit Comparison

	1-D GSE	1-D CST	1-D sequential analysis	2-D
No. of Examinees	8053	8053	8053	8053
No. of Estimated Parameters	47	67	114	113
Deviance	434507.78	566450.38	1000958.16	996530.83

GSE PLVL	GSE MC	GSE OE1	GSE OE2	CST MC	Logit	GSE Histogram	CST Histogram
					3.5		<
					3		<
							<
					2.5		<
							<
							<
							<
					2		<
						<	<
							X
						<	X
						<	X
	17				1.5	<	X
						<	XX
						<	XX
						X	XXX
						X	XXX
6	19		42.3		1	X	XXXX
				29 35		XX	XXX
		41.3		16 57		XXXX	XXX
5	24					XXXXX	XXXX
		41.2	42.2			XXXXXXX	XXXXX
	20			49	0.5	XXXXXXX	XXXXXXX
4	39			58 64		XXXXXXXXX	XXXXXX
	16 40			22		XXXXXXXXX	XXXXXX
	4		42.1	7 56		XXXXXXXXXX	XXXXXXX
	5 12 14			11 14		XXXXXXXXX	XXXXXXX
	13 28 37	41.1		18 21 41	0	XXXXXXXXXXX	XXXXXX
	11 18			15 37 43 44 54		XXXXXXXXX	XXXXXXX
	10 32			5 13 62		XXXXXXX	XXXXXX
	15 30			9		XXXXXXXXX	XXXXXX
	2 8 31			32		XXXXXXX	XXXXXXX
3	1 9 27 33 35 36			24 39 50 60	-0.5	XXXXXXX	XXXXXX
	21 29			63 65		XXXXX	XXXXX
	3 6 7			40 61		XXXXXX	XXXXX
				4 33 47		XXX	XXXXXX
	26 34			19 28 38 42		XXX	XXXX
2				10	-1	XX	XX
	38			17 26		X	XXX
	25			59		XX	XX
	22			2		<	XX
				25 31 45		<	X
				51 52	-1.5	<	X
				6 46		<	X
				27 30			X
							<
				1 3			<
				12 23 48	-2		<
							<
				8 34			<
				20			X
				36 55			
	23				-2.5		<
							<
							<
					-3		
							X
					-3.5		
					-4		
					-4.5		
1				53			

Note: "x" and "<" in GSE and CST histograms represent 50 and under 50 respectively.
Figure 9. 2-D calibration.

population parameter structure used in the present study. Compared to the deviance in the 2-D analysis (996530.83), the 1-D sequential analysis showed larger deviance (100958.19), indicating that the 2-D approach fitted the data better.

Summary and Discussion

With a common set of people, calibrations and linking were done to investigate the measurement properties of the GSE math and the CST math tests. We chose a sequential unidimensional approach to set up a common scale, followed by a 2-D approach. The latter was done with the anchored population structure based upon the previous sequential unidimensional approach to put the 2-D scales onto the common scale.

In summary, the sequential unidimensional approach results were very similar to the 2-D approach results. More than half of the GSE MC items (about 68% of GSE MC items) were not optimal with respect to the core purpose of GSE test, and they overlapped with the CST items in the middle and lower ability range. The GSE OE items appear to serve the test purpose very well; they differentiate between honors students and those that are not honors-eligible, and also within the honors students. To reduce testing time, as suggested in the results section, a shorter GSE tests (e.g., the two GSE OE items and perhaps some of the difficult MC items) could be combined with all or part of the CST items. These suggestions were supported by the item and test information analyses.

Did we truly achieve comparability through these procedures? Perhaps a more reasonable question would be how much successful the linking was. The answer depends on what level of comparability is desired and how much similarity between the linked tests exists. Recently several researchers offered the evaluation perspective on the quality of linking and its categorization. (Feuer, Holland, Green, Bertenthal, and Hemphill, 1999; Linn, 1993; Kolen, 2004; Kolen and Brennan, 2004; Mislevy, 1992). According to the most recent view on linking by Kolen et al. (2004), the test commonalities decide the quality of the linking and the degree of the commonalities can be examined in terms of test score inferences, test constructs, examinee populations, and measurement conditions. The first criterion is about the similarity of the use of test scores. The second criterion is about the amount of shared test dimensionality. The third criterion is about the target examinee population for tests. The last criterion is about the same test specification such as test length, test format, and test administration condition. To the extent that the GSE and the CST share the common features in these respects, the strength of the linking will be decided. The CST and GSE are not the same test. They have different test blueprints; The GSE is for relatively low stakes voluntary honors student selection; and the CST is a mandatory high stakes exam. Yet, it is reasonable to think that they share the same general curriculum: the three years of high school mathematics material in Algebra, Algebra II, and Geometry. This is a more broad definition than what is exactly measured by each test, but it provides a ground in our common scale construction for comparability—through the linking that uses the same framework with different test specifications (Feuer et al., 1999). Given this commonality and the estimated dimension correlation (0.743) from the 2-D analysis, the construct that the two test measure and are linked on, could be defined as the high level of education outcome that is high school math achievement.

One can build a common scale or the same metric following the procedure used in this study or use other IRT model estimate transformation methods (see, e.g., Kolen and Brennan, 2004). This gives different tests the same origin and unit of measurement, but it should be remembered that the quality of comparability will be decided by many aspects of the linking, which are the extent of the model-data fit, linking item invariance, the degrees of similarity in testing purposes, the constructs tests measure, examinee populations, and test specification.

References

Adams, R. J., and Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard and M.

Wilson (Eds.), *Objective measurement: Theory into practice* (Vol 3, pp. 143-166). Norwood, NJ: Ablex.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Adams, R. J., Wilson, M., and Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Adams, R. J., Wilson, M., and Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression, *Journal of Educational and Behavioral Statistics, 22,* 47-76.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., and Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3-26.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage.

Kolen, M. J. (2004). Linking assessments: Concept and history. *Applied Psychological Methods, 28*, 219-226.

Kolen M. J., and Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Linacre, J. M. (1994). *Many-facet Rasch measurement.* Chicago: MESA Press.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83-102.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 49-174.

Masters, G. N. (1988). The analysis of partial credit scoring. *Applied Measurement in Education, 1,* 279-297.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359-381.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS Policy information Center.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

von Davier, M., and von Davier, A. A. (2004). *A unified approach to IRT scale linkage and scale transformations* (ETS research report RR-04-09). Princeton, NJ: Educational Testing Service.

Wilson M. (1992). The ordered partition model: an extension of the partial credit model. *Applied Psychological Measurement, 16,* 309-325.

Wang, W. C. (1999). Direct estimation of correlations among latent traits within IRT Framework. *Methods of Psychological Research Online, 4,* 47-70.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Wu, M. L., Adams, R. J., and Wilson, M. R. (1998). *ConQuest: Generalized item response modeling software.* Camberwell, VIC, Austrailia: Australian Council for Educational Research.