Listening. Learning. Leading.®

Research Report
ETS RR–12-11

Why the Major Field Test in Business Does Not Report Subscores: Reliability and Construct Validity Evidence

Guangming Ling

June 2012

Why the Major Field Test in Business Does Not Report Subscores: Reliability and Construct Validity Evidence

Guangming Ling
ETS, Princeton, New Jersey

June 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editors: Daniel R. Eignor and James E. Carlson

Technical Reviewers: Frank Rijmen and Donald A. Rock

Abstract

To assess the value of individual students' subscores on the Major Field Test in Business (MFT Business), I examined the test's internal structure with factor analysis and structural equation model methods, and analyzed the subscore reliabilities using the augmented scores method. Analyses of the internal structure suggested that the MFT Business measures a unidimensional construct, which does not support reporting individual students' subscores. Augmented score analyses revealed that the observed total score could approximate the true subscore more accurately than the observed subscores, which do not support reporting individuals' subscores either. The results from these two approaches provided consistent evidence in support of the current practice of not reporting individuals' subscores on the MFT Business. The relationship between the two approaches is discussed, followed by a recommendation of an alternate method for future research.

Key words**:** subscore, augmented score, internal structure, reliability

Acknowledgments

An earlier version of this manuscript was presented at the annual convention of the National Council of Measurement in Education, at San Diego, California, in April 2009. The author thanks Bethanne Mowery and Jill van den Heuvel for their support and help when this study began. Thanks also go to Elizabeth Gehrig and Linda DeLauro for their help in editing an earlier draft of this report and to John W. Young, Cathy Wendler, Kathi Perlove, and Kathy O'Neill for their review and valuable comments. However, the sole responsibility for the opinions expressed in this report remains with the author.

Table of Contents

Page

List of Tables

Page

List of Figures

Page

Background

Scores on the subscales (or subdomains) of a test may provide students and test score users with more specific and detailed information on individuals' performance than the total score alone provides. This type of score becomes more important when a test measures multiple content areas or multiple underlying constructs. For example, a college exit test for undergraduate business majors may contain items related to different aspects of business knowledge or skills, such as accounting, economics, management, quantitative and information systems, and so on. Students, teachers, and parents may be interested in knowing the competence levels reflected in each of these aspects, in addition to the total test scores. A score can be derived from a subset of items that cover a particular aspect of business knowledge (or a subset of items that measure a common subconstruct), which is often called a *subscore* (ETS, 2000; see also Ferrara & DeMauro, 2007).

In recent years, there have been increasing demands for the reporting of subscores, especially individual students' subscores. Goodman and Hambleton (2004) reviewed the standardized tests in 11 U.S. states and two Canadian provinces, as well as three U.S. testing companies, and found they all provided certain information at the subdomain level (i.e., individual students' subscores). Haladyna and Kramer (2005) also found that subscores of different content areas, subdomains, and so on, are in great demand by stakeholders, regardless of the original purpose of the test when it was first developed.

This trend of increasing interest in subscores leads to a series of questions regarding the value of subscores, for example, whether subscores should be reported and what type of evidence is needed to support such practices. These questions have resulted in a broad discussion among different stakeholders and different score users (Ferrara & DeMauro, 2007). On one hand, having the subscores available and reported would make the test more appealing in the market by satisfying the strong demand from teachers, parents, and students. On the other hand, test developers and psychometricians may insist that subscores be prepared and examined carefully (as are the total scores) before being reported or being used in a particular context. From this perspective, empirical and theoretical evidence may be deemed necessary in support of subscore reporting practices.

Several steps would be required during the test development procedure to facilitate the reporting of individual students' subscores (ETS, 2000; Ferrara & DeMauro, 2007). For

example, it is necessary to ensure that each content or domain area, if it is intended to report its subscore, is well represented with equal or comparable numbers of items and sound psychometric properties (e.g., acceptable reliability; ETS, 2000).

When a test is not originally designed to report individual students' subscores, a post hoc evaluation is necessary if it is later decided to report subscores. Different methods and perspectives may be considered in these cases. As it is for the total score, high reliability is always required for a subscore. For example, Ferrara and DeMauro (2007) recommended that a subscore be reported if it has a high reliability (i.e., an internal consistency reliability estimate of .85 or higher) and does not correlate highly with other subscores; a subscore with a low reliability should not be reported.

However, as reliability is a statistic sensitive to test length (number of items; see Nunnally & Bernstein, 1994), the reliability of an observed subscore is often lower than that of the total test. Researchers have developed several methods to improve subscore reliability using statistical models. For example, Yen (1987) suggested an *objective performance index* (OPI) to combine information from all items in the test to estimate the true score on a subset of items from the same objective (subject, content, or domain). Wainer et al. (2001) developed a similar method, called the *augmented scores method*, which borrows information from all subscores to estimate the true score of a given subscale. The main difference between these two methods is that the OPI is based on the proportion of correct scores, while the augmented scores method is based on the observed number of correct scores.

The test's internal structure needs to be considered in addition to subscores' reliabilities. As Messick (1989, p. 43) pointed out, reported scores should be reasonably consistent with what is known about the structural relations of the construct in question (see also Loevinger, 1957). The inter-item relationship structure should reflect the nature and dimensions of the construct or domain, which should also be captured at the level of test scores and their interpretations (Messick, 1989, p. 43). The internal structure of the test could provide information related to what content or construct subareas are implied in the test design, how they are measured, and what the interrelationships are; it could determine if all test items or a subgroup of them should be summarized and reported in the form of a score (Peak, 1953). Understanding the internal structure would help to interpret the test scores as well as the subscores. For example, a single total score often implies a unidimensional construct or structure of the test, while a combination

of subscores and composite scores implies a hierarchical structure or multidimensional structure (Messick, 1989, p. 44).

The subscore's reliability and the test's internal structure both seem necessary when considering reporting subscores for a test. However, little empirical study can be found in the literature that has addressed or investigated the relationship between them. It is necessary to compare these two approaches and to see if they could provide consistent answers to the question related to subscore reporting. Three research questions are explored in this study, using the Major Field Test in Business (MFT Business) as an example:

1. Should the subscores be reported based on the augmented score analysis? Or more specifically, do individuals' observed content-based subscores approximate the true subscores more accurately than the augmented scores, for example, from the observed total score?
2. What is the internal structure of the MFT Business? Does it support the reporting of individual students' content-related subscores?
3. Do the answers provided by the two methods agree with one another?

Methods

Instrument and Data

MFT Business (ETS, 2008) is a comprehensive outcomes assessment of basic, critical knowledge obtained by students in a business major (e.g., for a bachelor's degree). Students take the Major Field Test after they successfully complete the major-required courses, typically in the last year of college study. The test includes 120 multiple-choice items in total. These items cover seven subdomains of business knowledge and skills. The seven subdomains or content areas are: (a) S1 - accounting, (b) S2 - economics, (c) S3 - management, (d) S4 - quantitative business analysis and information systems, (e) S5 - finance, (f) S6 - marketing, and (g) S7 - legal and social environment. The number of items varies from 12 to 21 across the seven subscales.

Two types of scores are currently reported for the MFT Business: individuals' total test scores and the aggregate-level subscores—the assessment indicators (AIs; see ETS, 2008). The total test scores of individuals are scaled between 120 and 200 from the observed number of correct scores. *Assessment indicators* are the average scores (on a scale of 20 to 100) in a particular content subdomain for a group of students, equivalent to the average subscore of all

students in a class, in a program, or of an institution for a particular content subdomain (ETS, 2008). The aggregate-level subscores indicate the mastery level of business knowledge for students who are nested in a group (i.e., as a class or a program) instead of the mastery level for an individual student. The MFT Business presently does not report individual students' subscores.

The data came from the MFT Business administered between 2002 and 2006, including a total of 155,235 students who took the test (the same form) during this time period and provided a response to every item. Each individual item was scored as 1 if correct and as 0 if incorrect. The total number of correct responses in the test is counted as the total score. Similarly, the number of correct responses in each subscale is counted as the subscore. Table 1 displays the descriptive statistics for the total test and each subscale, using number-correct scores.

Table 1

Subscales and Relevant Statistics

	Subscale							
	Account. (S1)	Econ. (S2)	Mgmt. (S3)	Quant. (S4)	Finance (S5)	Mktg. (S6)	Legal (S7)	Total test
Number of items	21	20	19	19	13	14	12	118
Mean	9.37	8.57	10.95	10.79	4.75	6.48	6.02	56.92
SD	3.50	3.22	3.30	3.27	2.28	2.27	2.16	14.98
Corr. with total score	.78	.78	.78	.81	.69	.66	.67	—
Cronbach's α	.64	.60	.64	.65	.53	.44	.43	.89
Expected reliability[a]	.57	.55	.54	.54	.45	.47	.43	

Note. Account. = accounting; Econ. = economics; Mgmt. = management; Quant. = quantitative business analysis and information systems; Mktg. = marketing; Legal = legal and social environment; Corr. = correlation.

[a]The expected reliability is computed based on the Spearman–Brown prophecy formula (Wainer et al., 2001).

The Management (S3) subscale had higher mean scores (10.95 out of 19 items) than the others, while the Finance subscale (S5) had lower mean scores (4.75 out of 13 items). The total test had a mean score of 56.92 out of 118 items, which means on average, students had slightly less than half of all items correct. The total test had a desirable reliability (Cronbach's α) of .89.

Methods and Analyses

Wainer et al. (2001) suggested using augmented scores to improve a subscore's reliability, which borrows from other subscores in the test to compute a particular subscore for individual students. More recently, Haberman (2005, 2007; Haberman, Sinharay, & Puhan, 2006; Sinharay, Haberman, & Puhan, 2007) found that reliability-based analysis is helpful in informing the decision of reporting subscores. His approach compares the root mean square error (RMSE) of the true subscore when it is predicted (or approximated) by the observed subscale score, the observed total test score, and the two scores conjointly. The rationale is that if the RMSE of the predicted true subscore using the observed subscores is smaller than the RMSE approximated through other scores or estimations, then the observed subscore contributes unique and substantial information beyond that provided by the total score (Haberman, 2005). In other words, the true subscores could be approximated more accurately from the observed subscores than from the other scores. Wainer et al.'s augmented scores method shares the same flavor as the methods suggested by Haberman and his colleagues. Both methods focus on borrowing information from other sections or subscales; they are also noted as the augmented scores method in this report.

In addition to the augmentation approach, Wainer et al. (2001) applied the Spearman–Brown formula to compute the expected reliability for each subscale, which they used to make inference about the test dimensionality. More specifically, the expected reliability is computed using the number of items in the subscale (*m*) and the total number of items in the test (*M*) in the Spearman–Brown formula, with the total test reliability coefficient (Cronbach's α, represented as ρ in the formula). The expected reliability of a given subscale (ρ_s) can be expressed as follows:

$$\rho_s = \frac{\rho}{\rho + (1-\rho) * \frac{M}{m}}$$

Furthermore, Wainer et al. (2001) suggested that when the expected reliability coefficient is "about equal to" (p. 357) the observed reliability (Cronbach's α) for each subscale, a unidimensional structure is likely to be supported. Otherwise, unidimensionality may be less likely, and multidimensionality is likely to be supported (see also Nunnally & Bernstein, 1994).

Descriptive analyses of item scores, subscores, and total test scores were conducted in the present study. Moreover, the reliability of each subscale, the total test, the correlations between subscores, and the correlations between each subscore and the total test score were analyzed and compared. Expected reliability was also computed for each subscale using the Spearman–Brown formula. Following the approach of Haberman (2005) and his colleagues, a set of RMSEs was computed for each of the seven subscales while using different observed scores to approximate the related true subscores.

It should be noted that the augmented scores approach fails to take into account the internal structure of the test. In many cases, the test's internal structure may influence the interpretation of the subscore and affect the decision of whether to report individual students' subscores (Messick, 1989). Analyzing the internal structure (or dimensionality) of the test and subscales provides additional evidence beyond that provided by the reliability analysis.

Two types of exploratory analysis were performed to evaluate the internal structure of the test based on the seven subscores. Both principal component analysis (PCA; Widaman, 2007) and exploratory factor analysis (EFA) based on the maximum likelihood method were performed using the variance–covariance matrix of the seven subscale scores. The SAS Factor Procedures (PROC Factor; SAS Institute, 2002) were applied here. Confirmatory factor analysis (CFA) was then performed using LISREL 8 (Jöreskog & Sörbom, 1999), based on the models selected from the exploratory analysis.

A common goal of structural equation modeling (SEM) is to approximate or reproduce the observed variance–covariance matrix (for continuous variables) or the tetrachoric–polychoric correlation matrix (for categorical variables). More specifically, a model-implied variance–covariance (or correlation) matrix is specified to approximate the observed variance–covariance (or correlation) matrix, with a set of parameter estimates that could minimize the distance between the two matrixes (Jöreskog & Sörbom, 1999; Raykov & Marcoulidies, 2000).

Many different measures or fit indices have been discussed in the literature. Four fit indices were used to evaluate the model-data fit because they are typically reported in the

literature (e.g., Browne & Cudeck, 1993; Hu & Bentler, 1999; Hooper, Coughlan, & Mullen, 2008; Raykov & Marcoulidies, 2000; Yu & Muthén, 2001). More specifically, these authors suggest that a model with a root mean square error approximate (RMSEA) value below .08, a Tucker–Lewis index (TLI) value above .90, a comparative fit index (CFI) value above .90, and a standardized root mean residual (SRMR) value below .1 can be considered an acceptable fit; a model with an RMSEA value below .05, a TLI (and CFI) value above .95, and an SRMR value below .08 can be considered a good fit.

Item-level factor analysis—a method for categorical or binary variables—was applied as well. The traditional factor analysis based on continuous observed variables may lead to a biased estimation of the latent construct and other parameters for categorical or binary variables, especially when the distributions of categorical variables severely deviate from the normal distribution (e.g., Woods, 2002). Woods revealed that traditional factor analysis with binary–categorical item scores results in biased estimation of the standard errors, biased significance tests, overestimation of the number of factors, and underestimation of the factor loadings.

Two approaches have been developed for item-level factor analysis, the probit-link-based limited information factor analysis approach (e.g., Jöreskog & Sörbom, 1999), and the logit-link-based full information factor analysis approach (Bock & Aitkin, 1981). The former models (probit-link-based models) are more widely used under the framework of SEM, for example, in LISREL (Jöreskog, 2002; Jöreskog & Sörbom, 1999), in MPLUS (Muthén & Muthén, 1998), and in EQS (Bentler, 2001), and were adopted in this study.

Four models were constructed for the MFT Business, based on the binary item scores. More specifically, a probit linking function between each binary item score and the underlying latent variable was specified in the model. The observed variance–covariance matrix for continuous variables was replaced with the tetrachoric correlation matrix, estimated based on the binary scores. The diagonally weighted least squares (DWLS) estimation method through LISREL 8.8 (Jöreskog & Sörbom, 1999) was applied to fit the model to the data.

First, a measurement model was constructed for each subscale separately. All items of a content area (e.g., all accounting items) were set as indicators of a latent factor (e.g., accounting-related knowledge, skill, or the subscale latent factor). This model was used to examine how items within the same content domain perform when specified to measure a common latent

construct. The subscale-based measurement model is typically considered a priori for a more complex structural model based on these subscales (Raykov & Marcoulidies, 2000).

Second, a one-factor measurement model was constructed to see whether all the test items measured a single factor. Such a model having a good fit would support the argument that the test measures a single construct, and a single total score would be appropriate to represent the test performance.

Third, a correlated seven-factor model was constructed to incorporate all subscales and all items associated with each subscale. The seven subscale-related factors were assumed to be correlated with each other (see Figure 1). This model is more complex than the one-factor model. A good model fit would suggest that the seven content-based subscales measure correlated constructs. However, if these correlations are all very strong, for example, greater than .80 or .90, this may suggest unidimensionality (e.g., Bagozzi & Yi, 1992; Kline, 2005; Stricker & Rock, 2008).

The fourth model (the higher-order factor model) was assumed to have zero correlations specified among subscale factors but to have a second-order common factor explaining the subscale factors (see Figure 2). Numerically, the correlation between a pair of subscales is equal to the product of the two standardized regression coefficients in the higher-order model. However, if the higher-order regression coefficients (standardized) are all .9 or greater, the subscales can be explained by a common factor to a great extent, which would support the notion of a unidimensional test structure (e.g., Kline, 2005; Sawaki, Stricker, & Oranje, 2008).

A similar set of guidelines as described earlier in the subscore level CFA was applied to examine the model-data fit for these item-level factor models. To compare the latter three models based on the whole test, the correlated seven-factor model is less parsimonious than the higher-order factor model and with fewer parameters to estimate. The one-factor model was the simplest model with the greatest degree of freedom among the three models, which could support reporting only the total score. When the fit indices are comparably acceptable or good, simple models (more parsimonious) are often preferred over more complex models, though the latter models fit slightly better (Mulaik et al., 1989; Steiger, 2007; Tabachnick & Fidell, 2007).

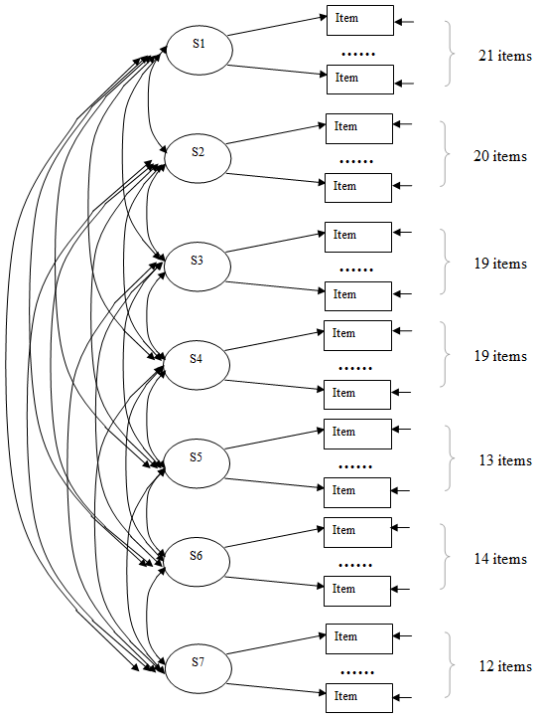

Figure 1. **The correlated seven-factor model using item scores.**

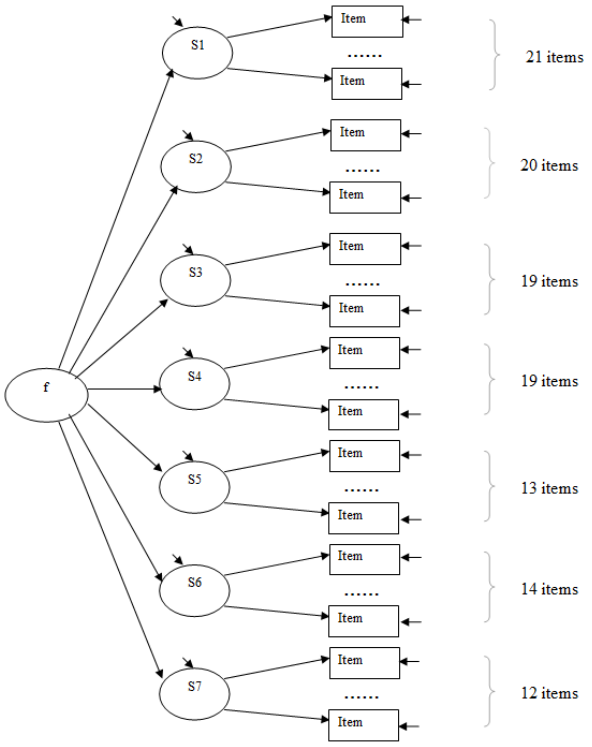

Figure 2. **The higher-order factor model using item scores.**

Results

Results of Descriptive and Reliability Analyses

Descriptive statistics showed that the reliability of the seven subscales ranged from .43 to .65 (Cronbach's α), much lower than the total test reliability (.89). These subscale reliabilities are considered low according to the commonly accepted minimum value of .70 in psychological and educational measurement (Nunnally & Bernstein, 1994, p. 265). The expected reliability of each subscale was also computed using the Spearman–Brown formula given the total test reliability (Nunnally & Bernstein, 1994; Spearman, 1904; Wainer et al., 2001). The expected reliability coefficient for each subscale was about equal to the observed reliability (Cronbach's α coefficient) as summarized in Table 1, with the average difference in the second decimal. Following the rationale of Wainer et al. (2001), this supports a unidimensional structure among the subscales.

Following the approach of Haberman and his colleagues (Haberman, 2005, 2007; Sinharay et al., 2007), four sets of RMSEs of the true subscore were computed and compared to the RMSE when it was estimated from the observed subscore, noted as SD(S_e). The four sets of RMSEs were computed using (a) the regression of the true subscore on the observed total score, SD(L-S_t); (b) the regression of the true subscore on the observed total score and the observed subscore, SD(M-S_t); (c) Kelly's estimation of the true subscore, SD(K-S_t); and (d) the approximated true error of the subscore, SD(F-D_t). The SD(S_e) value should have been smallest if the observed subscore could approximate or estimate the true subscore more accurately than the other scores (Haberman, 2005). However, results showed that the RMSE of the true subscale score estimated from the observed subscale score, SD(S_e), was consistently greater than those estimated from the observed total score, the combination of the subscale score and the total score, or the approximate true RMSE (see Table 2).

Using the first subscale, S1-Accounting, as an example, the SD(S_e) was 2.09, which represented the standard error of the true subscore of S1 when it was predicted from the observed subscore. The standard error for the observed total score, SD(L-S_t), was only 1.38, with the corresponding variance only less than half of the variance of S_e (see Table 2). This suggests that the approximation of subscale S1's true score from the observed subscore produced greater error than the approximation of the same true score of S1 from the observed total test score. Similarly,

the RMSEs of the other predictors or approximation methods were all smaller than that of the observed S1 subscore.

Table 2

Root Mean Square Error Comparisons Using Haberman's (2005) Approach

	Subscale						
	Account. (S1)	Econ. (S2)	Mgmt. (S3)	Quant. (S4)	Finance (S5)	Mktg. (S6)	Legal (S7)
No. of items	21	20	19	19	13	14	12
SD(S_e)	2.09	2.05	1.99	1.93	1.56	1.70	1.66
SD(K- S_t)	1.67	1.58	1.59	1.56	1.14	1.12	1.06
SD(L- S_t)	1.38	1.06	1.25	1.06	0.86	0.64	0.58
SD(M- S_t)	1.17	0.52	0.99	0.73	0.64	0.33	0.41
SD(F- D_t)	1.67	1.58	1.59	1.56	1.14	1.12	1.06

Note. SD(S_e) = observed subscore; SD(K- S_t) = Kelly's estimation of the true subscore; SD(L- S_t) = regression of the true subscore on the observed total score; SD(M- S_t) = regression of the true subscore on the observed total score and the observed subscore; SD(F-D_t) = approximated true error of the subscore; Account. = accounting; Econ. = economics; Mgmt. = management; Quant. = quantitative business analysis and information systems; Mktg. = marketing; Legal = legal and social environment.

The same patterns were found with the other six subscales (see Table 2), where the true subscore's RMSE was greater when it was predicted from the observed subscore than from the other predictors. These results suggest that the individual students' observed subscores could approximate the true subscores with greater errors than the total test scores or the combination of the total score and the subscore (Haberman, 2005).

Results of Subscale Score Factor Analysis

PCA and EFA were conducted based on the variance–covariance matrix of the seven observed subscores (number-correct raw scores) using PROC Factor (SAS Institute, 2002). The screeplot based on PCA is presented in Figure 3. A clear elbow appears at the second factor, with 54% of the total variance explained by the first factor and another 11% by the second factor. These results suggest the presence of one dimension (Cattell, 1966).

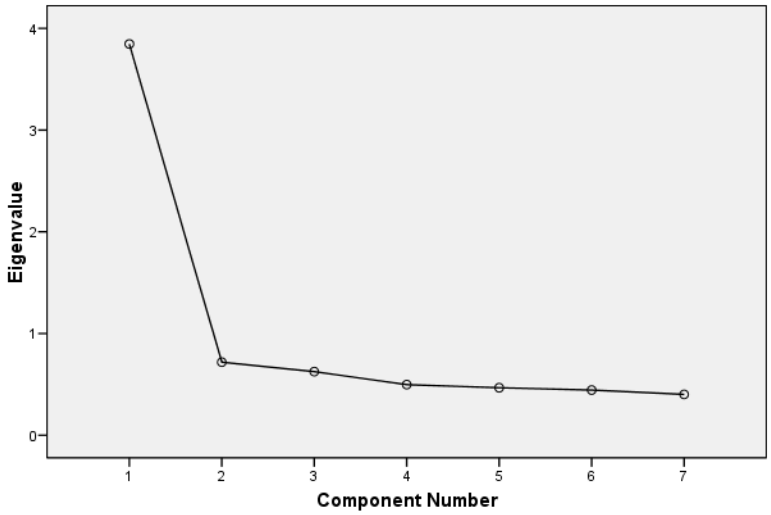

Figure 3. **Screeplot of eigenvalues from the principal component analysis.**

The one-factor model (see Figure 3) was then fitted to the data using LISREL 8.8 (Jöreskog & Sörbom, 1999). The fit indices suggest that the model fits the data very well (Hu & Bentler, 1999), with an RMSEA value of .064, a CFI value of .99, a TLI value of .98, and an SRMR value of .024. The standardized loadings were between .62 and .78. This well-fitted one-factor model supports the unidimensional nature of internal structure for the MFT Business.

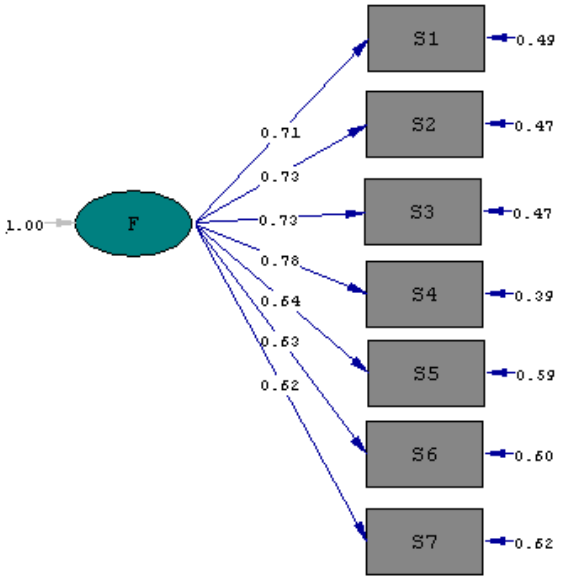

Figure 4. **Factor model and standardized estimates for the one-factor model using subscores.**

Results of Item-Level Factor Analyses

In the first stage, a measurement model was fitted to each of the seven subscales, where a single latent variable was indicated by all items belonging to a particular content subdomain. The measurement model fit the data well for each subscale. All of the RMSEAs were below the .05 criterion, and some of them were very close to .01; the CFIs and TLIs were all above .98 (see Table 3), suggesting that the distances between the model-implied correlation matrix and the tetrachoric correlation matrix were very small (Hu & Bentler, 1999).

Table 3

Model Fit Indices for the Subscales' Measurement Models Based on Item Scores

Subscale	RMSEA	90% CI RMSEA	TLI	CFI
S1-Accounting	.015	(.014–.017)	.99	.99
S2-Economics	.016	(.014–.017)	.99	.99
S3-Management	.015	(.013–.016)	.99	.99
S4-Quantitative Business Analysis and Information Systems	.020	(.019–.021)	.99	.99
S5-Finance	.016	(.014–.019)	.99	.99
S6-Marketing	.014	(.012–.016)	.98	.98
S7-Legal and Social Environment	.014	(.011–.016)	.99	.99

Note. RMSEA = root mean square error of approximation; CI = confidence interval; TLI = Tucker–Lewis index; CFI = comparative fit index.

Although item-factor loadings were all significantly different from zero (at the .05 level of significance), some were relatively low. For example, a total of 11 items had factor loadings below .20, including two items from S1-Accounting, three items from both S3-Economics and S7-Legal and Social Environment, and one item from each of S3-Management, S5-Finance, and S6-Marketing (see Table 4).

Table 4

Item-Factor Loadings of the Item-Level Factor Models

Subscale	Item	Subscale measurement model	One-factor model	Correlated seven-factor model	Higher-order factor model
S1	4	.14	.50	.51	.14
	7	.46	.07	.08	.39
	13	.42	.26	.27	.40
	24	.26	.37	.40	.19
	28	.36	.35	.37	.40
	31	.54	.26	.27	.59
	34	.19	.27	.29	.19
	38	.52	.19	.20	.54
	42	.35	.27	.29	.37
	48	.34	.39	.42	.32
	50	.40	.26	.29	.41
	82	.44	.50	.54	.53
	83	.50	.50	.55	.52
	86	.19	.35	.39	.16
	88	.25	.22	.23	.25
	95	.31	.39	.42	.21
	99	.37	.52	.54	.42
	108	.26	.20	.22	.29
	109	.34	.48	.52	.38
	110	.65	.28	.31	.56
	115	.37	.33	.35	.33
S2	5	.35	.22	.23	.35
	10	.49	.23	.23	.50
	11	.31	.27	.29	.27
	15	.50	.37	.39	.53
	19	.28	.30	.33	.21
	26	.22	.51	.53	.20
	51	.44	.51	.54	.44
	55	.35	.35	.38	.33
	56	.24	.34	.37	.22
	58	.50	.33	.36	.51

Subscale	Item	Subscale measurement model	One-factor model	Correlated seven-factor model	Higher-order factor model
S2	69	.48	.21	.23	.44
(cont.)	81	.59	.50	.52	.54
	85	.51	.42	.44	.47
	93	.39	.60	.62	.39
	94	.19	.35	.38	.16
	96	.34	.35	.38	.35
	103	.19	.45	.47	.16
	105	.28	.25	.26	.26
	107	.20	.48	.52	.23
	118	.17	.50	.51	.14
S3	16	.49	.28	.30	.50
	20	.32	.36	.37	.33
	22	.32	.16	.17	.29
	25	.46	.25	.28	.44
	27	.17	.47	.50	.14
	35	.42	.15	.17	.42
	39	.41	.13	.14	.38
	40	.24	.37	.40	.23
	46	.48	.42	.45	.47
	60	.38	.44	.47	.38
	70	.57	.32	.33	.53
	72	.31	.49	.51	.27
	77	.33	.49	.53	.38
	79	.40	.39	.42	.38
	84	.34	.13	.14	.32
	90	.46	.30	.31	.52
	111	.39	.34	.38	.39
	112	.27	.38	.39	.30
	117	.41	.48	.50	.47
S4	2	.21	.48	.50	.27
	6	.34	.16	.16	.32
	12	.36	.35	.38	.37

Subscale	Item	Subscale measurement model	One-factor model	Correlated seven-factor model	Higher-order factor model
S4	17	.40	.41	.44	.37
(cont.)	36	.29	.17	.19	.29
	43	.33	.36	.40	.33
	47	.43	.18	.19	.39
	57	.48	.32	.33	.51
	59	.36	.30	.33	.34
	65	.53	.38	.41	.52
	73	.26	.49	.51	.31
	76	.62	.36	.38	.62
	80	.53	.35	.39	.51
	87	.26	.49	.53	.26
	98	.52	.50	.51	.50
	100	.26	.49	.53	.28
	101	.51	.48	.52	.51
	114	.54	.19	.21	.54
	116	.39	.39	.41	.37
S5	14	.38	.52	.56	.27
	21	.38	.44	.47	.32
	30	.45	.20	.21	.40
	32	.07	.31	.33	.08
	37	.33	.41	.44	.29
	41	.35	.48	.51	.38
	61	.45	.38	.39	.54
	64	.45	.33	.35	.39
	67	.52	.13	.14	.55
	74	.39	.33	.34	.31
	75	.49	.31	.32	.40
	92	.41	.36	.40	.38
	106	.38	.23	.25	.55
S6	1	.11	.13	.14	.14
	3	.21	.28	.28	.29
	8	.22	.25	.26	.27

Subscale	Item	Subscale measurement model	One-factor model	Correlated seven-factor model	Higher-order factor model
S6	9	.24	.50	.55	.22
(cont.)	45	.25	.08	.08	.33
	49	.26	.28	.29	.22
	53	.27	.13	.14	.24
	54	.28	.21	.23	.23
	63	.28	.20	.22	.42
	66	.33	.42	.44	.54
	68	.35	.22	.24	.14
	71	.41	.21	.22	.23
	89	.52	.25	.27	.52
	91	.53	.28	.31	.31
S7	18	.17	.53	.54	.15
	23	.21	.35	.37	.17
	29	.54	.14	.15	.51
	33	.45	.29	.32	.45
	44	.32	.55	.59	.36
	52	.13	.36	.38	.14
	62	.35	.30	.32	.33
	78	.41	.30	.33	.42
	97	.46	.25	.27	.41
	102	.22	.30	.32	.23
	104	.05	.15	.16	.08
	113	.39	.16	.16	.43

The other three models—the one-factor model, the correlated seven-factor model, and the higher-order factor model—fit the data well, with acceptable to good fit indices (see Table 5). In the one-factor model, all 118 items were assumed to load on the single common factor, with all the residuals of these items assumed independent from each other (all covariances among residuals were set to 0; see Figure 1). The results suggest that this model had a good fit on most indices, with an RMSEA value of .015, an SRMR value of .025, a CFI value of .903, and a TLI value of .959 (see Table 5). The CFI value seems a bit lower than a good-fit criterion but still acceptable (Hu & Bentler, 1999).

Table 5

Model Fit Indices for the Three Item-Level Factor Analysis Models

	One-factor model	Correlated seven-factor model	Higher-order factor model
CFI	.903	.918	.932
TLI	.959	.965	.960
RMSEA	.015	.010	.010
SRMR	.025	.026	.026
df	6785	6764	6778

Note. CFI = comparative fit index; TLI = Tucker–Lewis index; SRMR = standardized root mean residual.

The item-factor loadings were all significantly different from zero, with some item loadings relatively low. A total of 17 items had loadings below .20 on the factor (see Table 5). A comparison with the previous results of the measurement model indicates that 15 out of the 17 items also had low loadings in the subscale measurement models. In summary, the fit indices of the second model suggest that the MFT Business measures a unidimensional construct.

The third model, the intercorrelated seven-factor model (see Figure 3), fit the data acceptably well by the CFI value of .918 but very well by the other fit indices (i.e., a TLI value of .965, an RMSEA value of .010, and an SRMR value of .026; see Table 5). The item-factor loadings were all significantly different from zero, with 16 items having a low item-factor loading (below .20). The fit indices, especially the CFI, suggest a slightly better fit than the one-factor model. In addition, the intercorrelations among the seven factors were very high, ranging from .83 to .95 (see Table 6), which indicates that the seven subscales are likely to measure a unidimensional construct.

Finally, the seven-factor model with a higher-order factor was fitted to the data. Instead of letting the seven subscale-related factors correlate with each other, a second-order common factor (*f*) was specified to be explained by the subscale-related factors (see Figure 2). The fit indices were very similar to the third model, the intercorrelated seven-factor model. The RMSEA value was .010, with an SRMR value of .026, a CFI value of .932, and a TLI value of .960 (see Table 5). The item-factor loading pattern was very similar to that in the correlated seven-factor model. The standardized regression coefficients between the second-order common factor and the seven subscale latent variables were all very high, ranging from .90 to .95 (see Table 6).

Table 6

Observed and Model-Implied Subscale Correlation Matrix

	Observed correlations between subscales and total test score						
	Account. (S1)	Econ. (S2)	Mgmt. (S3)	Quant. (S4)	Finance (S5)	Mktg. (S6)	Legal (S7)
S2	.53						
S3	.49	.52					
S4	.56	.56	.58				
S5	.52	.51	.41	.50			
S6	.39	.45	.49	.47	.36		
S7	.44	.43	.49	.48	.36	.38	
Total score	.78	.78	.78	.81	.69	.66	.67
Model-implied correlations between seven factors[a]							
S2	.91						
S3	.84	.89					
S4	.91	.93	.91				
S5	.93	.93	.83	.91			
S6	.83	.92	.95	.90	.83		
S7	.89	.91	.93	.93	.86	.90	
Standardized regression coefficients of seven factors on the higher order factor [b]							
	.90	.95	.91	.96	.90	.92	.94

Note. Calculations were based on the total test (118 items) and all valid observations. Account. = accounting; Econ. = economics; Mgmt. = management; Quant. = quantitative business analysis and information systems; Mktg. = marketing; Legal = legal and social environment.

[a] Estimated based on the correlated seven-factor model using item scores. [b] Estimated based on the higher-order factor model using item scores.

When comparing the latter three models, all had acceptable to good-fit indices. The fourth model—the higher-order factor model—fit the data as well as the correlated seven-factor model (the third model). Considering the high correlations estimated in the correlated seven-factor model and the higher-order factor model being more parsimonious, the latter one seems a better choice between the two models (Mulaik et al., 1989; Steiger, 2007). In addition, the extremely high values of the latent regression coefficient estimates (standardized) in the fourth model also suggest that the seven subscale-specific factors may likely measure a business-related general skill across domains of business knowledge and skills. The second model—the

one-factor model—fit slightly less well according to the CFI (slightly lower than the other two models) but fit the data comparably well by the other fit indices. In addition, the one-factor model had the fewest parameters among the three models. Combining the fit indices' and models' parsimoniousness, the one-factor model appears to be a better representation of the data, which also supports a unidimensional structure (Mulaik et al., 1989; Steiger, 2007; Tabachnick & Fidell, 2007).

Summary and Discussion

In summary, the results of this study suggest that the observed subscores of the MFT Business correlated moderately with each other and moderately to strongly with the observed total test score. Subscales had a much lower level of reliability than the total test in terms of Cronbach's α. The low reliabilities of the seven subscales and the very small differences between the expected and the observed reliability both indicate that is the test is most likely unidimensional (Wainer et al., 2001).

Analyses of reliabilities using the augmented scores approach showed that the RMSE of the subscale true score was greater when estimated by the observed subscore than by the total observed score or other predictors. In other words, for each subscale, the true score estimated by the subscale observed score was less reliable than the one estimated by the observed total test score. These results support the claim that reporting individual students' subscores may not add statistical information to the total score. These results also help answer the first research question: The content-based individual students' subscores did not add statistical information to the total score.

Traditional factor analysis using the subscores' variance and covariance information confirmed that a single-factor model fit the data well. The item-level factor analysis also demonstrated that a unidimensional structure could represent the MFT Business very well. These results suggest that although each subscale measured a content-specific subarea well, the MFT Business measures a unidimensional construct. This also answers the second research question: The MFT Business measures a unidimensional construct of business-related skills and knowledge and does not support reporting individual students' content-based subscores.

Combining the results from both augmented score analyses and the factor analysis of the internal structure, it supports a conclusion that the MFT Business should not provide individual students' subscores because such subscores do not add statistically reliable information over the

total test score. This also answers the third research question: The two approaches are consistent in supporting the current practice of not reporting individuals' content-based subscores for the MFT Business.

It should be noted that although the results of the two methods agree with each other in the context of the MFT Business, it may not necessary mean that such an agreement exists with other tests having different internal structure, subtotal correlations, and internal consistency.

The augmented scores method, or borrowing information from other subscales and the total test to estimate the true score of a given subscale, also imposes an implicit assumption about the unidimensionality of the test; that is, items of different subscales measure the same construct.

In the best scenario of a unidimensional test, the intersubscale correlations are typically very high and the subscales are less reliable than the total test; the observed subscores could provide some information about individual students' performance on the subscale but with less accuracy than that provided by the total test. In this situation, there is little theoretical or statistical evidence supporting subscore reporting, which is likely the case for the MFT Business in this study.

In a multidimensional test where each subscale measures a construct very distinct from the other subscales, the inter-subscale correlations will be relatively low. In these cases, the subscales' reliabilities will still be lower than that of the total test, but the differences may depend on the uniqueness of the subscales and the relationship between the subscale-specific factor and the common factor.

When using the augmented scores method, the reliability of a real test is estimated using Cronbach's α or other formulas based on the data, which typically provide a conservative approximation of the true reliability. As Lord and Novick (1968, pp. 47–50) pointed out, only in an essential tau-equivalent measurement or more restrictive model does the estimated reliability coefficient equal the true reliability. This may affect the results based on the reliability as most educational tests do not have the ideal feature of tau-equivalent measurement. Moreover, as suggested by Haberman (2005), the RMSE used in the augmented scores approach is determined by the reliabilities of the subscales and the total test (as well as the intersubscale correlations), which may confound the results as well. These two issues may require future research.

The analysis of a test's internal structure is a useful step to confirm the theoretical basis in terms of construct structure and could help to make decisions such as how to score the test and what scores to report (e.g., Messick, 1989). Using a factor analysis approach has added value to the augmented scores method because the former provides additional construct validity evidence of the test structure and links the subscores with the dimensionality of the test. A test with a multidimensional structure would strongly justify the meanings and interpretations of the subscores. However, more research is needed to evaluate if, in such cases, results obtained through the augmented scores and the internal structure approach agree.

Other methods, including the application of the SEM approach (e.g., Raykov, 2002), might also be considered for future research examining the reliabilities of subscores. Raykov's approach utilizes the measurement construct of the total test and the subscales, and provides an alternate estimate of the reliabilities of subscales. It would be interesting to investigate how these estimated reliabilities for the subscales differ from those estimated in the current study and the relationships among them.

References

Bagozzi, R. P., & Yi, Y. (1992). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16,* 74–94.

Bentler, P. M. (2001). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of the EM algorithm. *Psychometrika, 46,* 443–459.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Cattell, R. B. (Ed.). (1966). *Handbook of multivariate experimental psychology.* Chicago, IL: Rand McNally.

ETS. (2000). *ETS standards for quality and fairness*. Princeton, NJ: Author.

ETS. (2008). *MFT online tour.* Retrieved from http://www.ets.org/Media/Tests/MFT/demo/mftdemo_hi.html

Ferrara, S., & DeMauro, G. E. (2007). Standardized assessment of individual achievement in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 579–622). Westport, CT: Praeger.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17,* 145–220.

Haberman, S. J. (2005). *When can subscores have value?* (ETS Research Report No. RR-05-08). Princeton, NJ: ETS.

Haberman, S. J. (2007). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33,* 204–229.

Haberman, S. J., Sinharay, S., & Puhan, G. (2006). *Subscores for institutions* (ETS Research Report No. RR-06-13). Princeton, NJ: ETS.

Haladyna, T. M., & Kramer, G. (2005, April). *Poly-scoring of multiple-choice item responses in a high-stakes test.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, QC, Canada.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criterion versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Jöreskog, K. G. (2002). *Structural equation modeling with ordinal variables using LISREL*. Retrieved from http://www.ssicentral.com/lisrel/ordinal.htm

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL8: User's reference guide* (2nd ed.). Lincolnwood, IL: Scientific Software International.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennet, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105,* 430–445.

Muthén, L., & Muthén, B. O. (1998). *MPLUS: Statistical analysis with latent variables* [Computer software]. Los Angeles, CA: Authors.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Peak, H. (1953). Problems of objective observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 243–300). New York, NY: Dryden Press.

Raykov, T. (2002). Examining group differences in reliability of multi-component measuring instruments. *British Journal of Mathematical and Statistical Psychology*, *55,* 145–158.

Raykov, T., & Marcoulidies, G. A. (2000). *A first course in structural equation modeling.* Mahwah, NJ: Erlbaum.

SAS Institute. (2002). *SAS9 language reference: Dictionary, Volumes 1 and 2*. Cary, NC: Author.

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL®Internet-based test (iBT): Exploration in a field trial sample* (ETS Research Report No. TOEFLiBT-04). Princeton, NJ: ETS.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21–28.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42*, 893–898.

Stricker, L., & Rock, D. (2008). *Factor structure of the TOEFL Internet-based test across subgroups* (ETS Research Report No. TOEFLiBT-07). Princeton, NJ: ETS.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Allyn & Bacon.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., …Thissen, D. (2001). Augmented scores—"borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Hillsdale, NJ: Erlbaum.

Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors, and misconceptions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 177–203). Mahwah, NJ: Erlbaum.

Woods, C. M. (2002). Factor analysis of scales composed of binary items: Illustration with the Maudsley Obsessional Compulsive Inventory. *Journal of Psychopathology and Behavioral Assessment, 24*(4), 215–223.

Yen, W. M. (1987, June). *A Bayesian/IRT index of objective performance*. Paper presented at the annual meeting of the Psychometric Society, Montreal, QC, Canada.

Yu, C. Y., & Muthén, B. O. (2001). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes* (Technical report). Los Angeles: University of California at Los Angeles Graduate School of Education and Information Studies.

Notes

[1] In general, *subscore* refers to a score derived from a subset of items of a test. These items may measure a particular content, domain, or construct that is specified at the test development stage. It does not necessarily mean the score itself is reliable, nor does it mean the subscore should actually be reported. In this report, *subscore,* if not specified, stands for the subscore of individual students.

[2] The total test includes 120 items, but two items (Items 6 and 18 in the second section) were excluded from scoring. Thus those two items were excluded from this study as well.