ETS GRE®

Assess Ability. Predict Performance.

ETS GRE® Board Research Report
ETS GRE®–18-02
ETS RR–18-19

Statistical Properties of the *GRE®* Psychology Test Subscores

Yuming Liu

Frédéric Robin

Hanwook Yoo

Venessa Manna

December 2018

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

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

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board reports do not necessarily represent official Graduate Record Examinations Board position or policy.

The Graduate Record Examinations and ETS are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

****

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

Statistical Properties of the *GRE*® Psychology Test Subscores

Yuming Liu, Frédéric Robin, Hanwook Yoo, & Venessa Manna

Educational Testing Service, Princeton, NJ

The *GRE*® Psychology test is an achievement test that measures core knowledge in 12 content domains that represent the courses commonly offered at the undergraduate level. Currently, a total score and 2 subscores, experimental and social, are reported to test takers as well as graduate institutions. However, the American Psychological Association (APA) Board of Educational Affairs and the Council of University Directors of Clinical Psychology (CUDCP) have expressed a need for reporting new subscores. To meet such a demand, Educational Testing Service (ETS) content experts, in collaboration with the APA and CUDCP, proposed a new mapping of 12 content domains onto 6 subtest areas: biological; cognitive; social; developmental; clinical; and measurement, methodology, and other. In this study, we investigated the latent structure of the test and evaluated the statistical properties of the 6 subscores. Factor analyses showed that the test was essentially unidimensional. The disattenuated correlations among the 6 raw subscores were relatively high. However, when augmented with the total score, the subscores displayed low to moderate value added. Further, profile analyses showed that the augmented subscores displayed noticeably distinct profiles among test takers with the same total score. No evidence suggested that the content mapping of the 6 subscores was unreasonable. Additional validity research is needed to support the use of the 6 subscores, once their intended use is made clearer.

Keywords *GRE*® Psychology test; subscores; confirmatory factor analysis; reliability; score augmentation; value added

doi:10.1002/ets2.12206

The *GRE*® Psychology test currently reports a total score and two subscores: experimental and social. However the American Psychological Association (APA) Board of Educational Affairs and the Council of University Directors of Clinical Psychology (CUDCP) have expressed a need for new subscores that would provide an assessment of test takers' strengths and weaknesses and possibly contribute to the assessment of competencies in more narrowly defined undergraduate psychology content areas (Briel & Mills, 2014; Health Service Psychology Education Collaborative, 2013; Michel, 2012; M. Prinstein, personal communication, May 4, 2015). To meet such a demand GRE content experts in collaboration with APA and CUDCP staff mapped the 12 content domains that the test is designed to measure onto six subcontent areas: biological; cognitive; social; developmental; clinical; and measurement methodology and other. Preliminary analyses were conducted to explore the properties of these subscores (Kim, 2012; Michel, 2012; Puhan, Su, & Walker, 2012; Sinharay & Haberman, 2012; Walker, 2012). However these studies were limited in scope. In addition minor revisions to the test content were implemented in 2014 following the publication of the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*; American Psychiatric Association, 2013).

The purpose of this study is to provide evidence required for subscores beyond validity of content mapping conducted by the experts. According to Standard 1.14 of the 2014 edition of the *Standards for Educational and Psychological Testing* compiled by American Educational Research Association, American Psychological Association, and National Council for Measurement in Education (2014), such evidence would require that the subscores be sufficiently distinct from each other and be reliable to use in practice. To this end, we examined the properties of the six subscores using three test forms from the three most recent test administrations. We used confirmatory factor analyses (CFA) to assess whether the latent structure of the test is consistent with reporting a total score and six subscores. We then used classical test theory-based analyses to evaluate the reliability and value added of the subscores. In addition, we used profile analyses to evaluate whether the subscores are distinct from each other.

In the sections that follow, we first describe the content domains that the GRE Psychology test is designed to measure, the mapping of these domains onto the six subcontent areas, and the main findings of previous studies. We then describe the characteristics of the test data, including scoring and item screening procedures. In the Methods section, we first

Corresponding author: Y. Liu, E-mail: yliu@ets.org

describe the CFA used to assess the latent structure of the test and the appropriateness of the proposed content mapping. We then describe the classical test theory methods used to estimate subscore reliability and value added that the raw and augmented subscores could provide. We summarize our findings in the Results section. In the Discussion section, we present results of profile analyses and compare them with the results from other studies. Finally, in the Conclusions section, we summarize our findings and their implications for the development of the proposed subscores.

GRE Psychology Test

Current Test

The GRE Psychology test consists of 205 multiple choice items. Some items are self-contained, and others are related to the same stimulus materials, such as a description of an experiment or a graph. The test covers 12 broadly defined content domains that correspond to the fields of psychology taught at the undergraduate level (Educational Testing Service [ETS], 2014). Currently, two subscores that "enable assessment of strengths and weaknesses and can be used for guidance and placement purposes" (ETS, 2016a, p. 8) are reported in addition to a total score (Table 1). The content measured by the two subscores, experiment and social, was identified and approved by ETS test developers and an external test committee. Analyses showed that the reliabilities (Kuder–Richardson 20) were .89 and .88 for the experimental and social subscores, respectively, and that the disattenuated correlation was .89 between the two subtests, which justified the reporting of the two subscores (Briel, O'Neil, & Scheuneman, 1993).

Formula scoring has been implemented where correct responses are scored 1, incorrect −.25, and nonresponses (omitted or not attempted) 0. A test taker's raw total score is obtained by adding all the item scores, and his or her raw subscore is obtained by adding all the item scores in the subcontent. The raw total scores and subscores are first equated and scaled to account for form differences and then converted to final scores on the reporting scale (Kim, 2012; Puhan et al., 2012). The reporting total scores range from 200 to 990 in 10-point increments, and the reporting subscores range from 20 to 99 in one-point increments.

Proposed Subscores

The GRE Psychology test has been operating this way for a long time. However, there has always been an interest in developing more than two subscores, as indicated by McPeek, Altman, Wallmark, and Wingersky (1976). These authors proposed eight potential subscores and conducted reliability, correlation, and profile analyses based on a single test form of 50 items. They concluded that information for most test takers' strengths and weaknesses in some content areas could be obtained. However, it was determined that the subscores considered were not equally informative.

Recently there has been a renewed interest in developing additional subscores that could be used to evaluate students' strengths and weaknesses in more content areas than the two currently reported. Taking into account the needs of the APA and the CUDCP, GRE content experts identified six potential subscores that the current test could support (Michel, 2012; M. Prinstein, personal communication, May 4, 2015). Table 1 provides a detailed mapping of the 12 content domains to the six proposed subscores and the approximate number of items in each subcontent domain.

Because formula scoring makes the testing experience more complex to test takers and has been abandoned by almost all testing programs, it was decided that the GRE Psychology Test move to number-correct scoring. Therefore, number-correct scoring was used in the recent preliminary studies (Michel, 2012; Puhan et al., 2012; Sinharay & Haberman, 2012; Walker, 2012) as well as in this study.

Preliminary Findings

Puhan et al. (2012), Sinharay and Haberman (2012), and Walker (2012) examined the raw subscore distributions, correlations and reliabilities, and the extent to which individual subscore profiles differed across test takers. Based on data collected from several test forms administered in or prior to 2012, they found that the six subscore distributions were relatively well centered and normally shaped. This was expected, because there were 25–42 items associated with each subscore. The disattenuated correlations among the subscores were quite high (close to .9), and some of the subscores had reliability values lower than the desired value, .8, recommended for GRE subscore reporting (ETS, 2016a, p. 19; Sinharay,

Table 1 Mapping of GRE Psychology Content Domains to the Proposed and Current Subscores

Proposed subscore	Content domains	Current subscores	
Model I (6 factors)	Model IV (12 factors)	Model III (3 factors)	Model II (1 factor)
Biological (35 – 45 items)	Sensation and perception	Experimental psychology (approximately 80 items)	Total (205 items)
	Physiological and behavioral neuroscience		
Cognitive (35 – 50 items)	Learning		
	Language		
	Memory		
	Thinking		
Social (25 – 29 items)	Social	Social psychology (approximately 90 items)	
Developmental (25 – 29 items)	Lifespan development (childhood, adolescence, aging)		
Clinical (30 – 39 items)	Personality		
	Clinical and abnormal		
Measurement, methodology and other (31 – 39 items)	General	Other (approximately 35 items)	
	Measurement and methodology		

Puhan, & Haberman, 2011). However, they also found that by combining the total score with each subscore, the reliabilities of the augmented subscores were significantly increased. Such augmented subscores could provide added value beyond the total score, enhance subscore reliability, and increase the usefulness of individual subscore profiles. These findings helped us to better understand the statistical properties of the test and to select our assessment methods.

Test Data

Data from three recent test forms (Forms A, B, and C) developed in accordance with the *DSM-5* were available. Tests delivered in the United States (representing the majority of the test-taking population) were used, resulting in samples with 936, 939, and 1,072 valid test records for Forms A, B, and C, respectively. Generally, test takers had sufficient time to complete the test, as indicated by the small proportions of test takers who did not reach the end of the test and the relatively low ratios of not-reached variance to total score variance of .08, .02, and .04[1] for Forms A, B, and C, respectively.

Following standard postadministration procedures, item analyses and evaluations were conducted to remove poorly functioning items. As a result, the numbers of operational items were 205, 203, and 204 for Forms A, B, and C, respectively.[2] For each form, Appendix A shows the number of items, moments of score distributions (mean, standard deviation, skewness, and kurtosis) and reliability estimates (Cronbach's α coefficients) for the three forms and all models listed on Table 1. Overall, there are no apparent ceiling or flooring effects, subscores are approximately normally distributed, and results are consistent across the three forms.

Methods

Factor analysis, multidimensional item response theory (MIRT), and classical test theory approaches have been used to assess whether subscores have adequate psychometric quality (Sinharay et al., 2011). Factor analysis and MIRT are model-based methods that focus on describing the latent structure of a test. Although these methods may be used to estimate scores, they are difficult to implement operationally due to the relatively large sample sizes required to ensure estimation accuracy that may not be attainable by all the GRE Psychology test administrations. For these and other reasons,[3] it was decided to continue scoring the Psychology Test based on the observed raw test and subtest scores. In this context, the classical approach proposed by Haberman (2008), which does not require the fitting of statistical models, is particularly well suited for the estimation of the subscores and the assessment of their statistical properties.

In the following section, we describe the CFA conducted to assess the extent to which a 6-factor simple structure that corresponds to the contents underlying the six subscores fits the current test data. We then describe the analyses conducted to assess the reliability of each subscore and the extent to which each subscore may provide information beyond what

the total score has already provided (value added). Because some of the proposed subscores were insufficiently reliable on their own (Sinharay & Haberman, 2012), methods that combine a raw subscore with the total raw score (subscore augmentation) were also considered.

Confirmatory Factor Analyses

CFA is a well-established method that has been widely used in educational measurement as well as in many other disciplines (Brown, 2015; Finney & DiStefano, 2013). CFA is used to evaluate the extent to which a fully specified model fits the data. In this study, we assessed whether the 6-factor model of the proposed subscores fits the data well and whether three alternative models representing different content mappings would provide significantly different fit. Specifically, the following models were considered:

- Six-factor model representing the proposed six subscores.
- Single-factor model representing the total score.
- Three-factor model representing the current item classification: the two reported subscores and a category of the remaining items.
- Twelve-factor model representing the 12 content domains that the test is designed to measure.

A simple structure was specified for all the factor models; that is, each item contributes to one and only one factor (subscore), and all factors are correlated. The Mplus software (Muthén & Muthén, 1998) was employed to conduct the CFAs. Because the item response data were binary and the sample sizes were moderately large, we chose tetrachoric correlation for the estimation of the interitem correlations and weighted least square mean-adjusted (WLSM) estimator for the analyses.[4] The Mplus default scaling method was applied.[5] The thresholds indicate the *z*-scores that correspond to the probabilities of selecting the correct responses (Finney & DiStefano, 2013). The ratio of item threshold to factor loading represents item difficulty, whereas the ratio of factor loading to residual variance of the latent variable represents item discrimination.[6]

Because the model chi-square (χ^2) statistics are sensitive to sample size, we used the following statistics to evaluate model–data fit: the root mean square error of approximation (RMSEA; Brown, 2015, p. 71), the comparative fit index (CFI; Brown, 2015, p. 72), and the Tucker-Lewis index (TLI; Brown, 2015, p. 73). RMSEA is an index that puts a premium on model parsimony. CFI and TLI are indices designed to evaluate the fit of a specified model against the restricted "null" or "independent" model that fixes the covariances of all input variables to be zero. When dichotomous data are modeled, Brown (2015) and Yu (2002) have suggested that appropriate model fit is achieved when RMSEA values are close to or below .05 and CFI and TLI values are close to or above .95. According to Brown (pp. 74–75), methodologists generally agreed that CFI and TLI values below .90 should lead researchers to strongly reject the solution. CFI and TLI values in the range of .90–.95 may be an indication of acceptable model fit.

Reliability and Value Added of Raw and Augmented Subscores

It is important that the latent structure of a test is consistent with the proposed subscores and that each subscore is reliable and provides distinct information (American Educational Research Association et al., 2014). To assess the extent to which this statement holds true, Haberman (2008) and Sinharay, Haberman, and Puhan (2007) suggested a comparison of the reliability of an observed subscore to that of the observed total score as a predictor of the true subscore. They argued that if the observed subscore is a more reliable predictor of the true subscore than the observed total score, then the subscore provides *value added* over what has already been provided by the total score, which is considered distinct information.

To estimate the reliability of such predictors, Haberman (2005, 2008) proposed computing the proportional reduction of mean squared error (PRMSE[7]) of the predictor over that of the average observed subscore (a trivial predictor). He demonstrated that for each subscore, the PRMSE of the subscore is comparable to the subscore's reliability (Cronbach α coefficient). The PRMSE of the total score can be estimated given the estimates of the total score reliability, the subscore reliabilities and standard deviations, and the covariance among the subscores. In our analyses, we evaluated the value added of each subscore by comparing its reliability with its corresponding PRMSE of the total score.

As Sinharay et al. (2011) and Wainer et al. (2001) pointed out, it is often the case with educational tests that subscores are not reliable enough or do not provide enough value added on their own. Wainer and colleagues showed that when the

correlations among subscores are moderate to high, some of the information from the other subscores in the test can be used to *augment* a subscore. Haberman (2008) proposed a similar approach to subscore augmentation, in which the total score is used in combination with the subscore to produce the augmented subscore. For example, an augmented subscore for the biological subtest (H_B) may be obtained as follows:

$$H_B = \overline{B} + .46\left(\mathrm{B}_{\mathrm{raw}} - \overline{B}\right) + .10\left(T_{raw} - \overline{T}\right),$$

where $\mathrm{B}_{\mathrm{raw}}$ is a person's raw score on the biological subtest, $\overline{B}$ is the average subscore of the sample group, T_{raw} is the person's raw score on the whole test, and $\overline{T}$ is the average total score of the sample group.[8] Haberman showed that the weights (i.e., .46 and .10 in this example) can be determined based on the observed data.

Appendix B provides a brief description of the computation procedures to obtain the augmentation weights and the PRMSE values of augmented subscores. These statistics were estimated using an SAS version of the SQE software program developed by Yao, Sinharay, and Haberman (2014).

To interpret our results, we relied on the findings and recommendations suggested by Sinharay (2010) and Sinharay et al. (2011). The findings were based on data from several testing programs as well as simulated data under a broad range of simulation conditions that included test length, number of subscores, and subscore reliabilities and correlations. We also sought advice from Shelby Haberman, who suggested that an increase of .05 in PRMSE would indicate a moderate level of value added (S. Haberman, personal communication, January 4, 2016). In terms of subscore reliability, a desirable value for the GRE subject testing program is .80 or above (ETS, 2016a, p. 19). As with value added, commonly accepted guidelines for augmented subscore reliability are not available. Because worthwhile augmentation would require augmented subscore reliability to be higher than that of the raw subscore, we propose .85 as a minimum desirable value for implementation.

Results

Analyses were conducted on data collected from three operational test forms administered in the fall of 2014. By conducting analyses across different forms and test taker samples, the consistency of the results could be evaluated.

Confirmatory Factor Analyses

The fit statistics for the 6-, 1-, 3-, and 12-factor models are summarized in Table 2. RMSEAs ranged from .023 to .028, well below the .05 criterion recommended by Brown (2015) and Yu (2002). CFI and TLI values were greater than .95 for Form A and Form C, above the criterion recommended (Brown, 2015), and between .932 and .943 for Form B, slightly below the criterion, but still acceptable according to Brown (2015).

These results show that the construct being measured is essentially unidimensional and that all the factor models investigated provided a good fit to the data. The results were consistent with those based on a previous version of the test (McPeek et al., 1976) and supported the reporting of a total score. However, the results also showed that the 6-factor model representative of the proposed content mapping does not result in a significant loss in model fit when compared to the 12-factor model representative of the content areas the test is designed to measure.

Correlation, Reliability, and Value Added

Table 3 provides estimates of the correlations among the six raw subscores of interest. On average, Pearson correlations of the subscores ranged from .62 to .75, and the disattenuated correlations ranged from .77 to .92. The average disattenuated correlations of the cognitive and biological subtests tended to be the highest and the lowest, respectively, across all three forms.

As expected, disattenuated correlations were relatively high because the test is essentially unidimensional. However, based on analyses of data from other testing programs and simulated data, Sinharay (2010) showed that subscores at such high correlation levels can also have value added.

Table 2 Summary of Model–Data Fit Statistics of Confirmatory Factor Analyses

Model	χ^2	*df*	*p*-value	RMSEA			CFI	TLI
				Mean	Low	High		
Form A								
6-factor	28,736.12	18,899	.000	.024	.023	.024	.957	.956
1-factor	29,548.89	18,914	.000	.025	.024	.025	.953	.953
3-factor	29,128.17	18,911	.000	.024	.023	.025	.955	.955
12-factor	28,163.27	18,848	.000	.023	.022	.024	.959	.959
Form B								
6-factor	32,777.37	19,487	.000	.027	.026	.027	.939	.939
1-factor	34,315.42	19,502	.000	.028	.028	.029	.932	.932
3-factor	33,658.26	19,499	.000	.028	.027	.028	.935	.935
12-factor	31,891.24	19,436	.000	.026	.026	.027	.943	.942
Form C								
6-factor	30,682.37	18,705	.000	.024	.024	.025	.960	.959
1-factor	32,129.00	18,720	.000	.026	.025	.026	.955	.955
3-factor	31,367.59	18,717	.000	.025	.025	.026	.958	.957
12-factor	30,159.07	18,654	.000	.024	.023	.024	.961	.961

Note. RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index.

Table 3 Pearson (Above Diagonal) and Disattenuated (Below Diagonal) Correlations Among the Subscores and Average Correlations of Each Subscore With the Other Subscores

Form	Biological	Cognitive	Social	Developmental	Clinical	Meas/Meth/Other	Average	
							ObsCorr	DisattCorr
Form A								
Biological	—	.78	.59	.64	.68	.73	.68	.84
Cognitive	.93	—	.70	.71	.73	.81	.75	.92
Social	.75	.89	—	.65	.66	.67	.66	.85
Developmental	.83	.92	.90	—	.69	.68	.68	.89
Clinical	.83	.88	.86	.91	—	.73	.70	.87
Meas/Meth/Other	.88	.97	.86	.89	.89	—	.72	.90
Form B								
Biological	—	.71	.55	.61	.61	.63	.62	.77
Cognitive	.84	—	.73	.72	.72	.74	.72	.88
Social	.70	.91	—	.63	.65	.65	.64	.83
Developmental	.78	.91	.85	—	.66	.61	.64	.84
Clinical	.74	.87	.83	.86	—	.62	.65	.81
Meas/Meth/Other	.78	.90	.84	.80	.77	—	.65	.82
Form C								
Biological	—	.77	.66	.62	.68	.67	.68	.82
Cognitive	.90	—	.77	.69	.72	.76	.74	.90
Social	.80	.94	—	.66	.68	.71	.70	.88
Developmental	.79	.89	.89	—	.66	.65	.66	.86
Clinical	.81	.86	.86	.87	—	.69	.69	.85
Meas/Meth/Other	.80	.91	.88	.85	.85	—	.70	.86

Note. ObsCorr. = observed Pearson correlation coefficient; DisattCorr = disattenuated correlation coefficient; Meas/Meth/Other = measurement, methodology, and other.

Table 4 provides estimates of Cronbach's α coefficients, PRMSEs based on the total score and the Haberman augmented subscore for each of the six proposed subtests of the three forms. Value added was reported as the difference between the reliability (Cronbach's alpha) of the raw subscore or Haberman augmented subscore and the PRMSE of the subscore estimated based on the total score.

Results show that the raw social and developmental subscores did not reach the minimum desired reliability of .80 and that only the raw biological subscore provided value added by itself in Forms B and C. However, with score augmentation, reliabilities or PRMSEs of all the subscores increased to .85 or above. The greatest increase was on social and developmental

Table 4 Reliabilities Based on the Raw (Cronbach's α Coefficient), Total, and Augmented Subscores (PRMSE) and Value Added of the Raw and Augmented Subscores

			PRMSE		Value added	
Form	#Items	Cronbach's α coefficient	Total score based	Haberman augmented	Raw score	Haberman augmented
Form A						
Biological	41	.83	.83	.89	.00	.06
Cognitive	40	.84	.93	.94	−.08	.01
Social	26	.73	.79	.85	−.06	.06
Developmental	25	.71	.86	.88	−.15	.02
Clinical	35	.82	.84	.89	−.03	.05
Meas/Meth/Other	38	.82	.90	.92	−.08	.02
Form B						
Biological	37	.83	.75	.87	.08	.12
Cognitive	44	.86	.92	.93	−.06	.01
Social	25	.75	.80	.86	−.05	.06
Developmental	27	.73	.82	.86	−.09	.04
Clinical	35	.80	.79	.87	.01	.08
Meas/Meth/Other	35	.80	.81	.87	.00	.07
Form C						
Biological	40	.86	.82	.90	.05	.08
Cognitive	42	.85	.92	.93	−.07	.01
Social	28	.78	.86	.89	−.08	.03
Developmental	25	.71	.82	.86	−.11	.04
Clinical	34	.81	.82	.88	−.01	.06
Meas/Meth/Other	35	.82	.84	.89	−.02	.05

Note. Meas/Meth/Other = measurement, methodology, and other; PRMSE = proportional reduction of mean squared error.

subscores. All six augmented subscores resulted in value added, although the added values were inconsistent across forms. For example, the value added was below .05 for cognitive; developmental; and measurement, methodology, and other in one or more test forms.

Discussion

Researchers have cautioned about the risk of losing potential validity when subscore correlations become high due to augmentation. Skorupski and Carvajal (2010) and Stone, Ye, Zhu, and Lane (2010) showed practical examples demonstrating that, when the augmented subscore correlations were essentially 1, the augmented subscore profiles were nearly identical. For the GRE Psychology test and the tests investigated by the above researchers, Table 5 provides a summary of augmented subscore reliabilities and augmented subscore correlations.[9] Although augmentation noticeably increased the psychology subscore reliabilities and correlations, the effects were not as large as those of the aforementioned studies. Unlike the nearly equivalent score profiles showed by Skorupski and Carvajal, we found noticeably distinct augmented subscore profiles among the GRE Psychology test takers with the same total score. Figures 1 and 2 display the profile patterns for two sets of five randomly selected test takers (who took Form B) at two ability levels: an average ability level and a relatively high ability level.[10]

Therefore, we believe that the proposed content mapping can support the production of six reliable augmented subscores. However, additional validity research is needed to support the use of the proposed subscores, once the intended use of the subscores is made clearer by the requesting organizations. Potential directions for future research include further pursuing the preliminary subscore profile analyses conducted by Puhan et al. (2012) to identify distinct subscore profiles and to assess their prevalence in the data. We also suggest collecting external performance data (the corresponding undergraduate course grades and possibly faculty evaluations) in the six subcontent areas to supplement the existing data. Contrasting group analyses of such data to be conducted in collaboration with the requesting organizations would allow for performance standards to be developed (Cizek & Bunch, 2007).

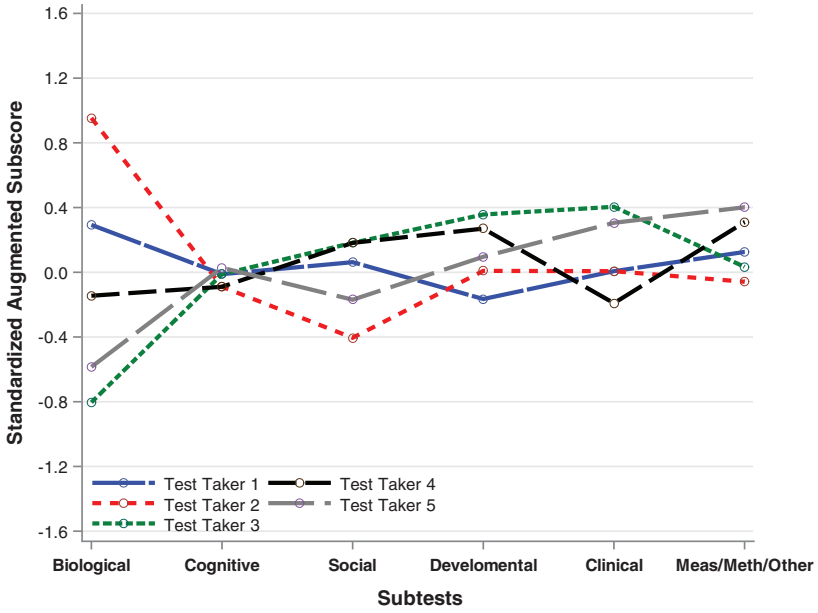

Figure 1 Standardized augmented subscore profiles of five randomly chosen test takers of an average ability level (standardized total score of 0).

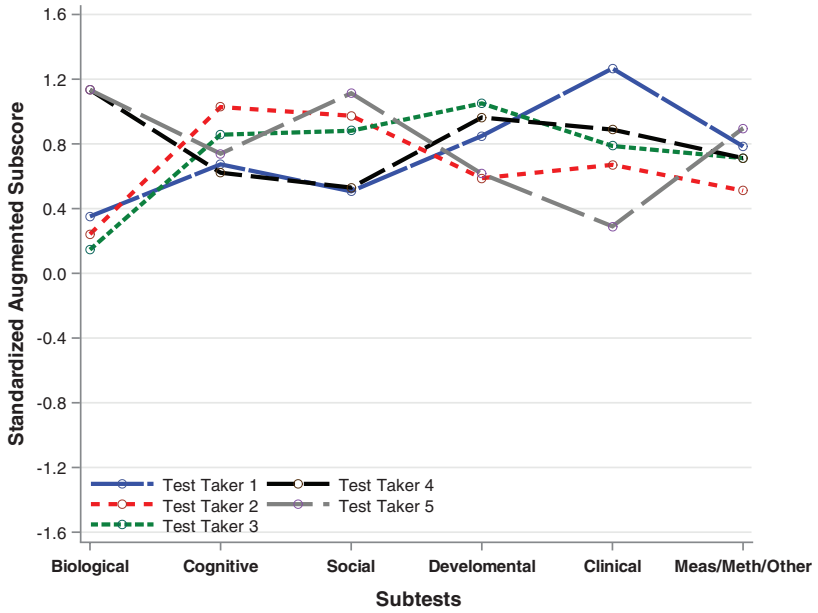

Figure 2 Standardized augmented subscore profiles of five randomly chosen test takers at a relatively high ability level (standardized total score of .76).

Table 5 Number of Subscores, Test Length, and Summary of Raw and Augmented Subscore Reliability and Correlation Statistics for Psychology and Other Tests

			Average reliability		Average correlation	
Form	# Subscore	Test length	Raw	Augmented	Raw	Augmented
Form A	6	205	.79	.89	.70	.95
Form B	6	203	.80	.88	.66	.92
Form C	6	204	.81	.89	.69	.94
Skorupski and Carvajal (2010)	4	64	.69	.89	.60 ~ .70	>.97
Stone et al. (2010)	4	59	.77	.93	.77	1.00

Conclusions

In this study, we evaluated the appropriateness of mapping the 12 content domains of the GRE Psychology test onto the six subcontent areas of interest and evaluated the statistical properties of the six corresponding subscores. Factor analyses showed that the test is unidimensional and that neither the 6-factor model nor the 12-factor model provide significantly better fit than the 1-factor model. However, no evidence showed that the content mapping of the six subscores identified by the content experts was unreasonable.

For the three forms analyzed, on average, the Pearson and disattenuated correlations among the six raw subscores were moderately high, but relatively low when compared to other assessments that report subscores (Sinharay, 2010). The raw subscore reliabilities ranged from .71 to .86, with those for social and developmental being the lowest. To increase these values, at least somewhat, test developers have agreed to slightly rebalance the number of items included in each subtest.

None of the raw subscores by themselves consistently produced value added over what has already provided by the total score. However, when used in combination with the raw total score, the augmented subscores displayed low (on cognitive; developmental; and measurement, methodology and other) to moderate (on biological and clinical) value added. The comparable reliabilities or PRMSE estimates of all the augmented subscores increased to .85 or above across the three test forms. These results indicate subscore augmentation is needed for the GRE Psychology test to reach the desired level of subscore reliability.

The profile analyses showed that, unlike tests with nearly equivalent score profiles discussed in other studies, the GRE Psychology test displayed noticeably distinct augmented subscore profiles among test takers with the same total score, indicating that GRE Psychology test subscores are distinct from each other. Further validity studies, including detailed profile analyses and development of performance standards, are needed to support the use of the proposed subscores.

Notes

1 As data were collected under formula scoring instructions, speededness was evaluated using the ratio of not-reached variance to total variance, with values above .15 indicating speededness (ETS, 2016b).
2 Further item screening was necessary for the CFA, as pairs of very easy or very hard items produced responses with a very high degree of similarity across test taker responses, which prevented the proper estimation of some of the interitem tetrachoric correlations. Thus, 196, 199, and 195 items were retained in the CFA for Forms A, B, and C, respectively.
3 Number correct raw score is simple to explain and easy to understand.
4 Following the suggestion of Dr. Finney (review comments were obtained on a previous version of this study), we explored both the WLSM and weighted least square mean variance-adjusted (WSLMV) estimation methods available in Mplus. In our preliminary analyses, the parameter estimates of WLSM and WLSMV were identical, but WLSMV took as many as 30 hours to converge, whereas WLSM took only 1 hour for a single data set.
5 Delta scaling and total variance of the underlying continuous variable for each item is set to 1.
6 With the delta or default scaling method in Mplus, item difficulty (b) and discrimination (a) of the two-parameter logistic item response theory (2PL IRT) model can be converted from threshold ($\boldsymbol{\tau}_D$) and factor loading ($\boldsymbol{\lambda}_D$). Specifically, $\boldsymbol{b} = \boldsymbol{\tau}_D/\boldsymbol{\lambda}_D$, and $\boldsymbol{a} = \boldsymbol{\lambda}_D/\sqrt{\boldsymbol{\Theta}_D}$, where $\boldsymbol{\Theta}_D$ represents residual variance of the latent variable (Finney & DiStefano, 2013).
7 PRMSE can be regarded as a subscore reliability statistic that is comparable to Cronbach's α coefficient (ETS, 2017, p. 30).
8 Note that because an augmented subscore is derived from the subscore and the total score, it is possible that test takers with the same raw subscore but different total scores will receive different augmented subscores. This situation is akin to that encountered when item response theory (IRT) pattern scoring is used.
9 Skorupski and Carvajal (2010) and Stone et al. (2010) used several alternative score augmentation approaches. However, to be able to make comparisons in Table 5, we only report the result they obtained using the same augmentation approach we used with the Psychology Test.
10 To facilitate interpretations, the scores were standardized to a mean of 0 and a standard deviation of 1. With augmented subscore reliability close to 0.9, variations in profile beyond $\pm$0.35 may be regarded as significant.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.
Briel, J. B., & Mills, C. (2014). *GRE psychology test* (Unpublished presentation). Princeton, NJ: Educational Testing Service.
Briel, J. B., O'Neil, K. A., & Scheuneman, J. D. (1993). *GRE technical manual*. Princeton, NJ: Educational Testing Service.
Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford.
Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications Ltd.
Educational Testing Service. (2014). *GRE® psychology test: Practice book*. Princeton, NJ: Educational Testing Service.
Educational Testing Service. (2016a). *GRE® guide to the use of scores 2016–17*. Princeton, NJ: Educational Testing Service.
Educational Testing Service. (2016b). *GRE test analysis report no. SR-2016-021* (Unpublished manuscript). Princeton, NJ: Educational Testing Service.
Educational Testing Service. (2017). *GRE® guide to the use of scores 2017–18*. Princeton, NJ: Educational Testing Service.
Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Charlotte, NC: IAP.
Haberman, S. (2005). *When can subscores have value?* (Research Report No. RR-05-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb01985.x
Haberman, S. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*, 204–229.
Health Service Psychology Education Collaborative. (2013). Professional psychology in health care services: A blueprint for education and training. *American Psychologist*, *68*, 411–426.
Kim, S. (2012). *Subscore equating study for GRE psychology test* (Unpublished memorandum). Princeton, NJ: Educational Testing Service.
McPeek, M., Altman, R., Wallmark, M., & Wingersky, B. (1976). *An investigation of the feasibility of obtaining additional subscores on the GRE Advanced Psychology Test* (GRE Board Professional Report No. 74–4P). Princeton, NJ: Educational Testing Service. Retrieved from http://files.eric.ed.gov/fulltext/ED163090.pdf
Michel, R. (2012). *Proposed subscore for the GRE psychology subject test* (Unpublished memorandum). Princeton, NJ: Educational Testing Service.
Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus (version 7.1) [computer software]*. Los Angeles, CA: Muthén & Muthén.
Puhan, G., Su, M., & Walker, M. (2012). *Subscore profile study for GRE psychology test* (Unpublished memorandum). Princeton, NJ: Educational Testing Service.
Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47,* 150–174.
Sinharay, S., & Haberman, S. (2012). *Evaluating the utility of augmented subscores for GRE psychology test* (Unpublished memorandum).
Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21–28.
Sinharay, S., Puhan, G., & Haberman, S. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30*(3), 29–40.
Skorupski, W. P., & Carvajal, J. (2010). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement*, *70*, 357–375.
Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education, 23*, 63–86.
Wainer, H., Vevea, J., Camacho, F., Reeve, 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). Mahwah, NJ: Erlbaum.
Walker, M. E. (2012). *Examinee performance on proposed subscore of the GRE psychology test* (Unpublished memorandum). Princeton, NJ: Educational Testing Service.
Yao, L., Sinharay, S., & Haberman, S. (2014). *SQE: A software package to evaluate the value of observed subscores and to produce subscores* (Research Memorandum No. RM-14-02). Princeton, NJ: Educational Testing Service.
Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Unpublished doctoral dissertation). University of California, Los Angeles, CA. Retrieved from https://www.statmodel.com/download/Yudissertation.pdf

Appendix A

Descriptive Statistics of Raw Scores by Content Mapping

		Form A						Form B						Form C					
Test	Subscore	#Items	Mean	*SD*	Skew	Kurt	Reliability	#Items	Mean	*SD*	Skew	Kurt	Reliability	#Items	Mean	*SD*	Skew	Kurt	Reliability
6-Subtest	Biological	41	23.79	6.33	−.33	−.34	.83	37	20.44	6.02	−.05	−.46	.83	40	22.77	7.00	−.08	−.52	.86
	Cognitive	40	26.30	6.37	−.49	−.07	.84	44	27.92	7.35	−.23	−.56	.86	42	27.02	6.88	−.38	−.28	.85
	Social	26	18.56	3.91	−.76	.52	.73	25	15.63	3.76	−.28	−.17	.75	28	18.34	4.34	−.48	−.04	.77
	Developmental	25	15.79	3.83	−.49	.16	.71	27	17.16	4.19	−.28	−.26	.73	25	16.50	3.75	−.31	−.07	.71
	Clinical	35	26.20	5.28	−.88	.49	.81	35	24.10	5.44	−.66	.10	.80	34	25.39	5.01	−.55	.08	.81
	Meas/Meth/Oth	38	24.18	5.92	−.35	−.42	.82	35	19.82	5.58	−.11	−.48	.80	35	24.52	5.46	−.66	.23	.82
Total Test	Total	205	134.83	27.55	−.53	.00	.96	203	125.07	27.52	−.20	−.48	.95	204	134.54	28.15	−.40	−.12	.96
3-Subtest	Experimental	81	50.09	11.98	−.40	−.20	.91	81	48.36	12.37	−.12	−.59	.91	82	49.79	13.06	−.22	−.37	.92
	Social	86	60.56	11.53	−.76	.42	.90	87	56.89	11.74	−.42	−.17	.90	87	60.23	11.58	−.46	.05	.90
	Other	38	24.18	5.92	−.35	−.42	.82	35	19.82	5.58	−.11	−.48	.80	35	24.52	5.46	−.66	.23	.82
12-Subtest	Physiological	27	17.67	4.47	−.49	−.18	.79	24	13.55	3.82	−.11	−.33	.75	27	16.18	4.92	−.19	−.47	.82
	Sensation	14	6.12	2.49	.12	−.41	.60	13	6.89	2.71	−.02	−.60	.66	13	6.60	2.62	.07	−.56	.64
	Language	7	3.12	1.71	.13	−.75	.55	6	4.15	1.55	−.58	−.50	.57	7	4.91	1.46	−.56	−.07	.48
	Learning	10	7.59	1.69	−.84	.65	.57	10	5.83	2.02	−.10	−.62	.58	9	5.89	1.92	−.44	−.35	.56
	Memory	14	9.59	2.56	−.44	−.22	.66	16	9.41	3.14	−.12	−.56	.71	15	8.22	2.99	.07	−.65	.70
	Thinking	9	6.00	1.85	−.55	−.10	.58	12	8.53	2.34	−.49	−.26	.64	11	8.01	2.02	−.94	.92	.62
	Social	26	18.56	3.91	−.76	.52	.73	25	15.63	3.76	−.28	−.17	.75	28	18.34	4.34	−48	−.04	.77
	Developmental	25	15.79	3.83	−.49	.16	.71	27	17.16	4.19	−.28	−.26	.73	25	16.50	3.75	−.31	−.07	.71
	Personality	10	7.54	1.89	−.79	.13	.61	10	6.09	2.09	−.29	−.47	.56	10	7.37	2.01	−.66	−.06	.66
	Clinical	25	18.66	3.89	−.90	.47	.76	25	18.01	3.96	−.79	.23	.75	24	18.03	3.59	−.54	−.03	.74
	Measurement	26	17.93	4.27	−.53	−.21	.78	26	16.72	4.69	−.28	−.57	.79	26	19.11	4.36	−.76	.34	.80
	General	12	6.25	2.45	−.10	−.51	.64	9	3.11	1.69	.46	−.08	.47	9	5.41	1.74	−.30	−.27	.49

Appendix B

Estimation Methods of Proportional Reduction of Mean Squared Error (PRMSE)

Following is a brief description of how computations of the PRMSEs may be conducted. For demonstrations of these formulas and detailed examples, see Haberman (2005, 2008) and Sinharay et al. (2011).

Let us consider a test's observed subscore s and observed total score x, and their corresponding subscore and total score-based predictors of the true subscore s_t:

$$s_s = \bar{s} + \boldsymbol{\rho}^2\left(s_t, s\right)\left(s - \bar{s}\right),$$

$$s_x = \bar{s} + \frac{\boldsymbol{\sigma}\left(s_t\right)}{\boldsymbol{\sigma}\left(x\right)}\boldsymbol{\rho}\left(x_t, x\right)\boldsymbol{\rho}\left(s_t, x_t\right)\left(x - \bar{x}\right),$$

where $\boldsymbol{\sigma}$ denotes standard deviation, $\boldsymbol{\rho}$ is the correlation, and $\boldsymbol{\rho}^2(s_t, s)$ is the sample subscore reliability.

It can be shown that

$$PRMSE\left(s_s\right) = \frac{E\left(E\left(s\right) - s_t\right)^2 - E\left(s_s - s_t\right)^2}{E\left(E\left(s\right) - s_t\right)^2} = \boldsymbol{\rho}^2\left(s_t, s\right)$$

$$PRMSE\left(s_x\right) = \frac{E\left(E\left(s\right) - s_t\right)^2 - E\left(s_x - s_t\right)^2}{E\left(E\left(s\right) - s_t\right)^2} = \boldsymbol{\rho}^2\left(s_t, x_t\right)\boldsymbol{\rho}^2\left(x_t, x\right)$$

where

$\boldsymbol{\rho}^2(x_t, x)$= total score reliability

$\boldsymbol{\rho}^2\left(s_t, x_t\right) = \frac{\left[\mathrm{Cov}(s_t, x_t)\right]^2}{Var(s_t)\,.\,Var(x_t)}$, which can be obtained given the sample subscore reliability, standard deviation, and correlation values, and because (a) the covariance between true subscores is the same as the covariance between the corresponding observed scores, and (b) the variance of a true score is the variance of the corresponding observed score multiplied by its reliability.

Now consider the augmented subscore, s_{aug}, based predictor of the true subscore s_t:

$$s_{aug} = \bar{s} + a\left(x - \bar{x}\right) + b\left(s - \bar{s}\right),$$

where

$$a = \frac{\boldsymbol{\sigma}\left(s\right)}{\boldsymbol{\sigma}\left(x\right)}\boldsymbol{\rho}\left(s_t, s\right)\boldsymbol{\tau},$$

$$b = \boldsymbol{\rho}\left(s_t, s\right)\left[\boldsymbol{\rho}\left(s_t, s\right)\right] - \boldsymbol{\rho}\left(s, x\right)\boldsymbol{\tau},$$

and

$$\boldsymbol{\tau} = \frac{\boldsymbol{\rho}\left(x_t, x\right)\boldsymbol{\rho}\left(s_t, x_t\right) - \boldsymbol{\rho}\left(s, x\right)\boldsymbol{\rho}\left(s_t, s\right)}{1 - \boldsymbol{\rho}^2\left(s, x\right)},$$

Then, it can be shown that

$$PRMSE\left(s_{aug}\right) = 1 - \left[1 - \boldsymbol{\rho}^2\left(s, s_t\right)\right]\left[1 - \boldsymbol{\rho}^2\left(x, s_t.s\right)\right],$$

where $\boldsymbol{\rho}^2(s, s_t)$ is the reliability estimate of subscore s, and $\boldsymbol{\rho}^2(x, s_t.s)$ is the squared partial correlation of the observed total score and true subscore given the observed subscore.

Suggested citation:

Liu, Y., Robin, F., Yoo, H., & Manna, V. (2018). *Statistical properties of the GRE® Psychology test subscores* (GRE Board Research Report No. GRE-18-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12206

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

ETS, the ETS logo, GRE, and the GRE logo are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/