Developing Form Assembly Specifications for Exams With Multiple Choice and Constructed Response Items

Balancing reliability and validity concerns

Amy Hendrickson
Brian Patterson
Maureen Ewing

The College Board

Paper presented at the Annual Conference of the National Council for Measurement in Education, May 1, 201, Denver

Developing Form Assembly Specifications for Exams with Multiple Choice and Constructed Response Items Sections

The psychometric considerations and challenges associated with including constructed response items on tests are discussed along with how these issues affect the form assembly specifications for mixed-format exams. Reliability and validity, security and fairness, pretesting, content and skills coverage, test length and timing, weights, statistical specifications, and equating concerns are all discussed. Practical suggestions for dealing with these concerns in the form assembly specifications are presented. Advanced Placement Program® (AP®) Exams are used as examples throughout, as most all of these large-scale standardized exams contain both multiple-choice and constructed response items.

Introduction

Test specifications, or form assembly specifications, specify the important attributes of items and test forms. They are blueprints that provide direction for test form construction. The test specifications should flow from the test purpose and use and include the distribution of content and skills, as well as statistical specifications across the form (Schmeiser and Welch, 2006). Documented and detailed test specifications are essential in any large-scale testing program to ensure that assessment developers build test forms that are consistent over time, with the goal of maintaining the comparability of scores across forms.

Schmeiser and Welch (2006) indicate that one begins the specification process by defining the “explicit linkage between test purpose and the criterion that defines the test domain, that is, what will and will not be measured in the test” (pg 308). The decision to include both multiple-choice (MC) and constructed response (CR) items on an examination generally stems from the belief that these item types can measure different and important constructs in the test domain, such as assessing deeper understanding or the ability to organize an argument. For many exams, correlations between MC and CR scores show, however, that the MC and CR items are measuring highly similar constructs. This is true for most all Advanced Placement Program® (AP®) Exams with the exception of those in history, English, and the world languages. Still, the use of CR items are highly valued and these items are thought to afford the opportunity to “provide a more direct measure of student achievement than traditional multiple-choice items” (Lane & Stone, 2006, p.389) Thus, from this perspective, in order to make valid interpretations of scores from the examination, both item types must contribute to the scores. However, inclusion of CR items on a test presents psychometric challenges. Specifications for mixed-format exams, therefore, must be written such that test forms built to these specifications balance validity concerns, as well as produce scores with reasonable statistical quality.

Form assembly specifications (FAS) for any exam have several parts that must be considered and decided upon. They include the following: content and skill distribution, test length, statistical specifications, and equating requirements. For exams with both multiple-choice and constructed response, these decisions often need to be considered both separately and jointly for each item type.

The purpose of this paper is to review each of these and other FAS issues in the context of mixed-format tests, but the psychometric considerations and challenges associated with CR items alone are discussed first. Advanced Placement exams will be used as examples throughout as these large-scale standardized exams contain both MC and CR items.

Psychometric Considerations and Challenges Associated with Constructed Response Items

In the context of large-scale standardized assessments, CR items are more challenging to deal with than MC items. During the item development process, it is difficult to control the psychometric characteristics of the CR items including the difficulty, discrimination, and parallelism of the items, as well as the reliability of their scores. CR items require more time for students to complete and more time and money for to score, compared to MC items. When constructed response items are included on a test, the following psychometric properties or issues should be considered.

Reliability and Validity

Proponents of CR items state that, in order to make valid interpretations of scores from examinations of some constructs, both MC and CR item types must contribute appropriately to the scores, whether this means that the item types are equally weighted or one of the item types is weighted more heavily (Kane and Case, 2004). However, scores from CR items are generally less reliable than scores on multiple-choice items, and, thus, from a purely psychometric standpoint, it is best to give proportionally higher weight to the more reliable multiple-choice section. For example, Ewing, Huff, and Kaliski (2010) reviewed reliability results for AP science exams in 2007 and 2008 and reported that the internal consistency of the MC scores was generally very high (ranging from .86 to .94), while the coefficient alpha of CR scores were not as high (ranging from .73 to .90). They also reported high coefficient alphas for AP composite scores. This indicates that both the MC and CR items together are generally reliable but the CR scores alone would not be as reliable as the composite score or multiple-choice scores alone unless many more CR items were added.

One important psychometric issue for mixed-format exams is the correlation between the scores on the MC and CR items. On one hand, the purpose of including both items types is most often because it is believed that distinct skills underlie the construct of interest and are best measured by different item types, thus leading to a low correlation between scores from these item types but contributing to the validity of the scores from the test. On the other hand, a high correlation of scores on these two sections is desirable as this helps to increase the reliability of the total test scores and may contribute to other statistical characteristics of the test, depending on the weighting of the sections, how equating is conducted, etc.

Security and Fairness

It is psychometrically desirable to administer the same CR item on multiple operational forms either to get pretest statistics (discussed next) or as a common item for equating purposes (discussed later). Despite that, reuse of broad, easily-remembered CR items over multiple forms presents security and fairness concerns. Students may be able to memorize CR questions and “release them” to the larger examinee population (via the web, discussion with friends and family, etc.). If these items were reused on later exam forms and some students had seen the items previously, then they would have an unfair advantage over students who had not seen the items.

Pretesting

Ideally, all operational items on a test form should have been previously administered to obtain pretest statistics and to determine any issues with the items. This pretesting allows for elimination of and/or modifications to poorly performing items and for assembly of forms that meet desired statistical characteristics. Pretest statistics from larger samples of motivated examinees best approximate the item statistics estimated on the basis of a full, operational administration.

For MC items, pretesting is fairly easily accomplished and is best done by embedding pretest items within operational forms. For CR items, however, pretesting is more difficult due to the relatively greater student time required, higher scoring costs, lack of rater consistency over time, and greater threats to the security of the items. Yet, pretesting CR items is just as important as pretesting MC items, perhaps even more important, due to the unpredictable performance of CR items.

As previously mentioned, CR items require relatively more time for students to complete, thus embedding pretest CR items in an operational form will take away a substantial amount of students’ operational testing time and thus reduce the amount of information used to calculate their score for operational purposes.

Any item statistics that are calculated from pretesting may change when the items are operationally administered due to changes to the items themselves (wording, location within the test, time allotted, etc.), changes to the students (ability, opportunity to learn, etc), or to changes in the scoring (Brennan, 1992). These changes may have more of an impact for CR items, especially because they are subjectively scored. Not only will the student samples change from pretesting to operational testing, but the raters of the CR items are also likely to change. These shifting factors that may introduce relatively more bias in the estimation of the statistics than they do for MC items, all mean that, even if CR pretest statistics are calculated, they may not be as stable as those for MC items.

Mixed-format FAS Considerations and Challenges

The challenge for developing the form assembly specifications for mixed-format exams is not in specifying what is desired, but in specifying requirements that can be operationalized on forms of the exam. The following are some of the major areas that need to be considered in developing the FAS. Some of these areas lend themselves to and have been researched, some have not, but each is important in the development of the FAS.

Content and skills

A key question that is asked very early in the test development process is: What are the content and skills (i.e., the domain) to be measured by the test? There are well-documented approaches for determining the content and skills to be measured on achievement tests. Traditional methods include gathering input from subject-matter experts (SMEs) who serve on test development committees, to reviewing of textbooks and other curriculum documents (e.g., course syllabi, state and national standards) and finally to use of empirical results from large scale curriculum studies (Schmeiser & Welch, 2006). Methods for organizing and prioritizing content and skills have also been developed (Wiggins & McTighe, 2005) which can be particularly helpful for subject areas with large amounts of content to be addressed (e.g., history).

Once subject matter experts have articulated the content and skills needed to appropriately cover the domain, subject matter experts and test developers should jointly determine how the content and skills are to be distributed over the item types. A clear articulation of the content and skills such as through conducting a domain analysis and developing a domain model and associated task models with evidence-centered design can help to articulate/clarify and support this distribution (Ewing, Packman, Hamen, & Thurber, in press; Hendrickson, Huff, & Luecht, in press).

Some construct-relevant skills may naturally lend themselves to one of the item types, for example writing skills may be best assessed through a CR essay item. However, many skills and content areas can and should be assessed with both item types and measuring them using different methods is only likely to increase the accuracy with which we tap the underlying trait (Hendrickson, Huff, & Luecht, in press).

Test length, number of items per type, timing of each section

As with any form assembly specifications, the length and time of the test must be decided. This means the number of items, the time allowed for each item, and time for directions and other administration events (breaks, etc.). These numbers should be established to ensure adequate coverage of the domain of content and skills, but will most likely also be dictated by logistical constraints (for example, a maximum of 3 hours).

For mixed-format tests, the number of items and testing time per item type must also be decided. The challenges to determining the number of test items per type are that a certain number of CR items must be included to meet the validity concerns, but the CR items take longer to complete and contribute less reliable information to students' scores, compared to the MC items. To help determine the best test length and given sufficient pretest or operational testing data, the Spearman-Brown prophecy formula (Haertel, 2007) can be used to predict the reliability of scores based on different numbers of each type of item. As to testing time, it may be that even within a test form, one constructed response item would require relatively more time to complete, than other CR items.

Weights

Tests with both MC and CR item types often report composite scores that are usually linear combinations of the scores from each of the sections. The test developers must decide which linear combination to use for the item scores on the MC and CR items—which likely have different numbers of score points—when calculating the composite score for the test. This decision revolves around the desired amount of weight each component will contribute to the composite and the use of different weights impacts both the reliability and validity of the scores.

Kolen (2006) discusses how the use of some weights may lead to lower composite score reliability than the reliability of the MC score alone, because of the relatively low reliability of the CR items. Only when the CR and MC sections can be shown to be congeneric are the weights that maximize reliability equal to those that will maximize criterion-related validity, however (Penev and Raykov, 2006). In addition, use of weights that maximize reliability may decrease the emphasis on the skills required for the constructed response section and hence detract from face validity (Kane and Case, 2004). These findings emphasize that selection of the weights requires finding a balance between reliability of the composite and the desired relative contribution of each item type to the composite.

Weighting Schemes. Kolen (2006) reviewed various weighting strategies for mixed-format tests. The type of weights generally used that have been researched includes: relative weights, observed score effective weights, true score effective weights, weights based on the number of score points, weights chosen to maximize reliability, IRT weights, and validity weights, which require a criterion (Kolen, 2006; Wainer and Thissen, 1993).

For many mixed-format tests, including the Advanced Placement Exams, weights are assigned based on the desired proportion contribution of each item section to the total composite raw score. For AP Exams, what we call the relative operational weights are sometimes referred to as the nominal, logical, or *a priori* weights and are published in official descriptions of the exams. These relative weights may be termed "nominal" because they are based on professional judgment and not on any statistical methodology. For example, the AP 2006 English Literature and Composition Exam uses 0.55 and 0.45 as the relative operational weights for the CR and MC sections, respectively, which were decided upon by a test development committee composed of AP teachers and post-secondary faculty. Thus, the CR section contributes 55% of the total composite weighted score. Use of such *a priori* weighting schemes follows a construct-representation argument and ignores the statistical and psychometric characteristics of the tests, such as the correlation between the subscores on the two item types, the relative contribution of each item type to the total test, and the test score reliability.

Effective weights, on the other hand, provide an index of the contribution of variability for each item type to the composite. The effective weight can be interpreted as the proportion of composite score variance that is attributable to a component of the composite (Kolen, 2007). Effective weights can be determined in terms of observed scores or true scores, if using a psychometric model. Considering the AP 2006 English Literature and Composition Exam again, while the operational weight for the CR section is 0.55, the effective weight, that is, the proportion of total test variance accounted for by the CR section, is 0.46.

Moses, Dorans, Miao, and Yon (2006) proposed applying the same weight to each CR item as is applied to each multiple-choice item for the AP Exams, and referred to this as item-score weighting. They state that this scheme minimizes the influence of CR score unreliability, among other psychometric benefits.

Weights that maximize test score reliability can be determined, also (Gulliksen, 1950; Kolen, 2007; Walker, 2005). Wainer and Thissen (1993) found that for the AP European History Exam containing two sections of unequal reliability (0.90 for multiple-choice and 0.46 for constructed response, the median values over 1982-1986) the equal weighting of the sections resulted in a composite reliability of 0.80 versus an optimally weighted composite reliability of 0.90.

Weighting Research with AP. To help inform the choice of weights for the MC and CR sections of the AP exams, we have conducted research on the effect of various weights on the reliability and predictive validity of scores from the MC and CR sections. Examinees from the 2006 administrations of the AP exams in Biology and Spanish Literature, as well as from the 2005 and 2006 administrations of the English Language & Composition exam were identified from among those entering 110 colleges and universities in the fall of 2006.

Our reliability results indicate that use of different weighting schemes does impact the reliability of the scores for the AP Exams, especially for those with high CR section weights. The increase in reliability by using the maximum reliability weighting scheme was as much as 0.029, taking the reliability of the Spanish Literature exam from 0.856 to 0.885. By the Spearman-Brown prophecy formula (Haertel, 2007), increasing the reliability from 0.856 (the composite score reliability under the current weighting scheme) to 0.885 (the composite score reliability under the maximum reliability scheme) is equivalent to adding 19 parallel MC items. Thus, while this increase may seem small in terms of the reliability value, when this increase is translated into the number of MC items that would need to be added to cause this increase, the impact is more evident.

Analyses were completed using weights that maximize reliability and those that maximize predictive validity of the composite for predicting students' mean first-year college course grade in the relevant discipline. The analyses conducted were first to estimate coefficient alpha for a variety of different MC and CR weight sets that maintained the same range for the resulting composite. Next, the predictive-validity-optimizing weights were estimated by first running a random intercepts multi-level model of the following form:

$$Y_{ij} = \beta_{0j} + \beta_1 MC_{ij} + \beta_2 FR_{ij} + \varepsilon_{ij} \quad \text{with} \quad \varepsilon_{ij} \sim \mathrm{Normal}(0, \sigma_j^2)$$

where Y_{ij} is the first-year subject-area GPA of student i attending college or university j and MC_{ij} and FR_{ij} are that student's unweighted multiple-choice and constructed response section scores for the relevant AP Exam. Since the GPA variable (Y_{ij}) is not on the same scale as the operational composite after weighting and since the maximum composite is fixed at the operational composite maximum, the ratio of the parameter estimates for multiple-choice and constructed response ($\hat{\beta}_1$ and $\hat{\beta}_2$, respectively) leads us to the weights we need to linearly combine the section scores into a composite. In particular, the weight for MC that is used to construct the composite is given by:

$\omega_{MC} = [\hat{\beta}_1 \cdot \max(Composite)] / [\hat{\beta}_1 \cdot \max(MC) + \hat{\beta}_2 \cdot \max(FR)]$. Finally, partial F-tests were conducted to determine if the operational weights or the weights that maximize reliability were equal to the weights that optimize predictive validity.

Table 1 presents the relative weights, section and composite reliabilities and F-Tests for comparing the operational and maximal reliability weight ratios to those that maximize predictive validity. The ratio of MC and CR section weights that maximize the predictive validity of the composite for predicting students' mean first-year course grade in the relevant discipline is statistically significantly different from the ratio of current AP section weights for the Biology 2006 exam [F(1, 4287) = 29.468; p < 0.001], but not for the English Language 2005 ([F(1, 2184) = 0.202; p = 0.653] or 2006 exams [F(1, 3160) = 3.702; p = 0.054]. The operational relative weights differ from those that maximize predictive validity quite substantially for the Biology 2006 exam, less for the English Language 2005 exam, and least for the English Language 2006 exam. In all three cases, the weights that maximize predictive validity place more weight on the MC section than is currently applied. The predictive validity weights are most different from the operational weights for the Biology 2006 exam, for which the FR section had the highest reliability of the three FR sections (.819) and for which the MC/FR correlation was the highest among the three tests (.860).

These results indicate that the validity maximizing weights may be dependent on the statistical characteristics of the items, including the reliability, and that these weights may be different at each administration. The difference between the English Language 2005 and 2006 results may be explained in a couple of different ways. First this difference could be due to the form differences being confounded with differences across the samples. Since only a single cohort of entering freshmen were analyzed, all students who took the exam in 2005 were juniors in high school, while all of those who took the exam in 2006 were seniors. In other words, the difference could be caused by differences in the groups within our cohort. Second, the results could be due to a difference in how well each section predicts at either two or one year out from when the criterion is measured.

We also compared the ratio of MC and CR section weights that maximize the predictive validity to the ratio of weights that maximize the reliability of the composite. These ratios were found to be statistically significantly different for the Biology [F(1, 4287) = 21.206; p < 0.001] and English Language [F(1, 3160) = 8.1481; p = 0.004] 2006 exams, but not for the English Language 2005 [F(1, 2184) = 0.548; p = 0.4592] exam. The weights that maximize the reliability of the composite differ from those that maximize predictive validity quite substantially for the English Language 2006 exam, less for the Biology 2006 exam, and least for the English Language 2005 exam. The weights that maximize reliability are higher for the MC section than is currently applied for the Biology 2006 exam, but are lower for the English Language exams.

Figure 1 demonstrates the extent to which the reliability of the composite for the 2006 administration of the AP English Language and Composition varies with different constructed response and multiple-choice section weights. Plotted horizontally are the relative CR weights and the vertical reference lines designate the operational relative CR weight (dashed line) and the relative CR weight that optimizes predictive validity (solid line). Again, these findings show that the weights that maximize reliability are not necessarily the same weights that will maximize criterion-related validity.

Work by Swaminathan and Rogers with AP exams (this session) indicates that the relative weights of MC and CR sections also affects the misclassification rates of students into AP grade levels. Their recommendation is, that if relative weights are to be used, they should match the test design, or alternatively, the test design should be designed to reflect the desired weighting scheme.

These findings all illustrate the fact that the decision of how to combine item-type scores represents a trade-off between the test measuring what it is meant and expected to measure and the test measuring consistently (Walker, 2005). Selection of the weights requires finding a balance between reliability of the composite and the desired relative contribution of each item type to the composite.

Statistical Specifications

Form assembly specifications include statistical specifications for the items and test as a whole. These statistical specs are often stated in terms of difficulty and discrimination values within a classical test theory context (often p-value and point-biserial coefficients) and/or in terms of item information values from an item response theory model. These specifications may include specific item requirements (e.g., “the minimum item discrimination value should be 0.25”) and whole form distributional requirements (e.g., “the mean and standard deviation of multiple-choice item difficulties should be 0.50 and 0.10”). For mixed-format assessments, these statistical requirements may be specified for each item type.

Even minimal statistical specifications require that items have pretest item statistics. If only some or even none of the items are pretested, which is more likely for the CR items, then minimal to no statistical specifications may be required. The stronger the pretesting model, the tighter, more detailed the statistical requirements may be. That is, if all items have stable pretest statistics then distributional characteristics may be specified, for example, ‘10% of items must have a p-value between 0.40 and 0.50’.

Desired reliability values for each item type section may be stated also, but these specifications may not be able to be determined until after large-scale pretesting or a full form has been administered. Reliability considerations for the CR items could also include rater reliability specifications, such that the consistency of ratings is assessed and meets some minimal level.

Equating

Equating forms of a mixed-format test can be challenging in a number of ways. If the common-items nonequivalent groups design is used, the composition of the common items is of concern. Ideally, the common item set should be a mini version of the entire test in terms of content and skill distribution, test structure (i.e., item type), and statistical composition (Kolen and Brennan, 2005).

The common item set for a mixed-format exam, thus, should most likely include a representative set of both MC and CR items. This is challenging in two ways: 1. determining a representative set of CR items may be difficult, and 2. including the same CR items on multiple forms presents a security and fairness issue. As to the representativeness, because of the length of time required to complete these item types, relatively few CR items are generally included on a mixed-format assessment, thus there are few items to choose from for the common item set. Choosing a set that represents the content and skills assessed by all of the CR items from this small number of items may not be possible (Baxter et al., 1992; Dunbar et al., 1991; Haertel and Linn, 1996; Linn, 1995; Wainer, 1993b). The security and fairness concerns with reuse of CR items were discussed previously.

Further requirements for equating may include that the statistical properties of the equating common item set be parallel to those for both the old (reference) and the new forms and that the equating item set covers the full range of difficulty of the exam. If CR items are included in the common item set, then the statistical properties of all of the CR items should be represented by the common item set. This may be difficult to accomplish with only a few CR items and the statistical properties of the CR items/section may not be stable over multiple forms or administrations.

Suggestions

Although the inclusion of CR items on an examination requires more consideration and presents more challenges, many steps can be taken to increase the usefulness of these items on a test form:

1. Use a clearly articulated domain model.
2. Create shorter, clearly articulated CR items as this allows for inclusion of more CR items (due to less student time) of which a representative subset may more easily be identified, makes pretesting (especially embedded) more feasible, and allows for use of CR items for common item equating more feasible while minimizing security concerns.
3. Include relatively more of these CR items on each test form as this will likely increase the reliability of the CR section and improve the use of CR items for equating.
4. Assess all of the content and skills with MC **and** CR items as this may increase the correlation between scores on the MC and CR sections.
5. Pretesting
    a. Use embedded pretesting if possible, for at least the MC items. If CR embedded pretesting is not feasible, then use non-operational pretesting.
    b. Have at least some loose CR statistical specifications, in order to allow for the real possibility of not meeting those specifications but also to give test developers more tangible targets. The stronger the prestesting model, the tighter the statistical specifications may be. If strong pretesting (i.e., with larger samples of motivated examinees) is not possible for the CR items, then

looser CR statistical specifications should be stated and/or more tolerance for not achieving the specifications.

c. Pretest all item types or all task models, rather than all items. In this way, some information about how these types of items will perform is gathered, without having to pretest every item.

d. Require that only a sub-set of included items need to be pretested (e.g., 65% of items must be pretested).

6. Use CR weights that consider the statistical characteristics of the items and the overall test reliability.
7. Equating

a. If possible, use CR items for common item equating. Development of a constructed response items that are more amenable to inclusion in the common item set will make this more feasible. For example, the FAS may specify that relatively more, but shorter constructed response items be included, in order for a representative common item set to be more easily determined and used for equating forms.

b. Use of an MC-only common item set for the equating is most appropriate for exams with a high MC/CR correlation and may work well for exams with certain psychometric characteristics (see Lee et al, this session), but not for all mixed-format exams (DeMauro, 1992; Hanson, 1993).

Taking these steps may help to alleviate some of the concerns about and limitations of including CR items on assessments. Any test development process and the simultaneous specifications for forms assembly require a balance of reliability and validity concerns. This is especially true for mixed-format exams, but implementing some fairly straightforward requirements (e.g., weighting based on statistical characteristics or that reflect the test design) can go a long way to ensure that both reliability and validity aspects are achieved consistently across forms.

References

Baxter, G. P., Shavelson, R.J., Goldman, S.R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement*, 29, 1-17.

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, 5, 225-264.

Brennan, R. L. (2001). *Generalizabilty Theory*. New York: Springer.

DeMauro, G.E. (1992). *An investigation of the appropriateness of the TOEFL test as a matching variable to equate TWE topics* (Report 37). Princeton, NJ: Educational Testing Service.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education,* 4, 289-303.

Ewing, M., Huff, K., & Kaliski, Pamela (in press). Validating AP exam Scores: Current research and new directions. In P. Sadler, G. Sonnert, R. Tai,& K. Klopfenstein (Eds). Cambridge, MA: Harvard Education Press.

Ewing, M., Packman, S., Hamen, C., & Clark Thurber, A. (in press). Representing targets of measurement using ECD. *Applied Measurement in Education*.

Gulliksen (1950). *Theory of Mental Tests*. New York: Wiley.

Hanson, B.A. (1993). *A missing data approach to adjusting writing sample scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Haertel, E. H. (2006). Reliability. In R.L. Brennan (Ed.), *Educational Measurement*, 4th Edition (65-110), Washington, DC: American Council on Education.

Haertel, E.H., & Linn, R.L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Education Statistics.

Hendrickson, A., Huff, K., and Luecth, R. (in press). Claims, Evidence and Achievement Level Descriptors as a Foundation for Item Design and Test Specifications. *Applied Measurement in Education*.

Kane, M. & Case, S.M. (2004). The Reliability and Validity of Weighted Composite Scores. *Applied Measurement in Education*, 17(3), 221-240.

Kolen, M. & Brennan, R. (2004). *Test Equating, Scaling, and Linking: Methods and Practices.* (2nd ed.). New York: Springer-Verlag.

Kolen, M. J. (2006). Scaling and Norming. In R.L. Brennan (Ed.), *Educational Measurement, 4th Edition* (155-186), Washington, DC: American Council on Education.

Lane, S. & Stone, C. (2006). Performance Assessment. In R.L. Brennan (Ed.) *Educational Measurement, 4th Edition* (387-431). Washington, DC: American Council on Education.

Leighton, J. P. and Gierl, M. G. (Eds.), *Cognitive diagnostic assessment for education* (19 – 60). New York, NY: Cambridge University Press.

Linn, R. L. (1995). High-stakes uses of performance-based assessments: Rationale, examples, and problems of comparability. In. T.H.R.K. Oakland (Ed.) *International perspectives on academic assessment. Evaluation in education and human services* (pp. 49-73). Boston, MA: Kluwer Academic Publishers.

Moses, T. P., Dorans, N. J., Miao, J., & Yon, H. (2006). Weighting strategies for Advanced Placement exams (*Educational Testing Service Statistical Report SR-*

2006-31). Princeton, NJ: Author.

Penev, S. & Raykov, T. (2006). On the Relationship Between Maximal Reliability and Maximal Validity of Linear Composites. *Multivariate Behavioral Research*, 41(2), 105–126.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R.L. Brennan (Ed.) *Educational Measurement*, 4th Edition (307-353). Washington, DC: American Council on Education.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement*, *30*(1), 1-21.

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*, 103-118.

Walker, M. (2005). Forming weighted composites from multiple subscores (*Educational Testing Service Statistical Report SR-2005-41*). Princeton, NJ: Author.

Wiggins, G. & McTighe, J. (2005). *Understanding by design* (2nd Edition). Alexandria, VA: Association for Supervision and Curriculum Development.

Table 1. Relative Weights, Reliabilities and F-Tests for whether Weight Ratios Maximize Predictive Validity

			Reliability			**Relative Weight**			**DF**			
AP exam	**Year**	**Corr (MC, FR)**	**MC**	**FR**	**Composite**	**MC**	**FR**	**Weight Set**	**Num**	**Den**	**F**	**p**
English Language & Composition	2005	0.629	0.889	0.685	0.9726	0.4500	0.5500	Operational	1	2,184	0.20	0.6530
					0.9727	0.4175	0.5825	Maximum Reliability	1	2,184	0.55	0.4592
					0.9722	0.4968	0.5032	Maximum Predictive Validity	▨	▨	▨	▨
English Language & Composition	2006	0.598	0.879	0.678	0.9711	0.4500	0.5500	Operational	1	3,160	3.70	0.0544
					0.9715	0.3797	0.6203	Maximum Reliability	1	3,160	8.15	0.0043
					0.9681	0.5719	0.4281	Maximum Predictive Validity	▨	▨	▨	▨
Biology	2006	0.860	0.933	0.819	0.9567	0.6000	0.4000	Operational	1	4,287	29.47	0.0000
					0.9570	0.6412	0.3588	Maximum Reliability	1	4,287	21.21	0.0000
					0.9480	0.8361	0.1639	Maximum Predictive Validity	▨	▨	▨	▨

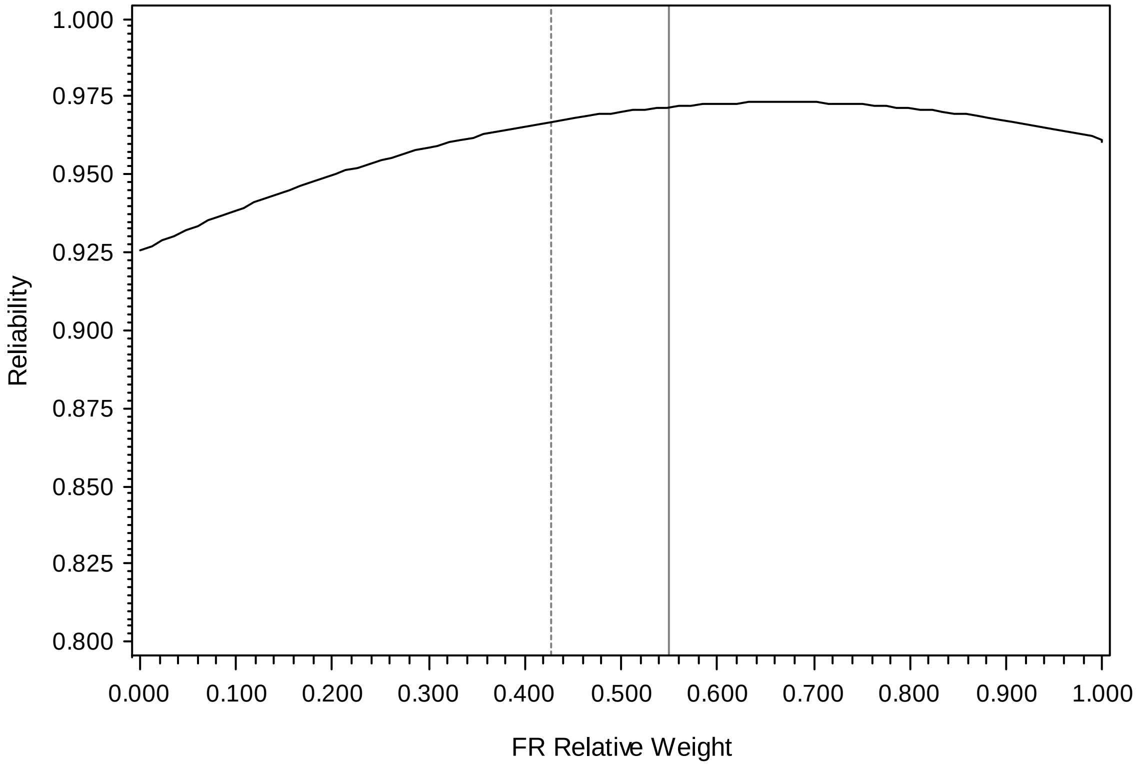

Figure 1. CR Reliability of English Language and Composition Composite with Varied Relative Weights