Extending a Practical Method for Developing Alternate Test Forms Using Independent Sets of Items

Frederick L. Oswald, Alyssa J. Friede, Neal Schmitt, Brian H. Kim and Lauren J. Ramsay

Organizational Research Methods 2005; 8; 149
DOI: 10.1177/1094428105275365

Extending a Practical Method for Developing Alternate Test Forms Using Independent Sets of Items

FREDERICK L. OSWALD
ALYSSA J. FRIEDE
NEAL SCHMITT
BRIAN H. KIM
LAUREN J. RAMSAY
Michigan State University

This study describes alternate test form development for a Situational Judgment Inventory (SJI) predicting college performance. Principal axis factor analysis of responses to the SJI lent support for a general factor, yet each SJI form sampled items across 12 distinct rationally derived content areas. The first step of developing alternate forms involved random and representative sampling of SJI items across each content area, creating a large number of preliminary 36-item SJI test forms. Gibson and Weiner (1998) provided criteria for selecting alternate forms; however, the authors of the present study extended this approach in the next step of selecting alternate forms based on their estimated criterion-related validity with grade point average. Results provide initial support for the 144 alternate forms generated. This general approach reflects a practical and methodologically sound means of developing alternate forms of types of measures that are rationally heterogeneous yet empirically homogeneous.

Keywords: *parallel forms; test banks; item banks; test development; reliability analysis*

In applied psychological and educational settings, standardized tests tend to be administered to groups of individuals in a variety of settings, where repeated test administrations take place over time. Therefore, the concern for test security is often high, and it may be desirable to have a large bank of test items from which to generate multiple forms of a test, whether forms are administered in a traditional paper-and-pencil format, in a computer-adaptive test (in which the length of the test is not necessarily

Authors' Note: The authors would like to thank the College Board for its support in conducting this research. Correspondence concerning this article should be directed to Frederick L. Oswald, Department of Psychology, Michigan State University, East Lansing, MI 48824-1117; e-mail: foswald@msu.edu.

Organizational Research Methods, Vol. 8 No. 2, April 2005 149-164
DOI: 10.1177/1094428105275365

fixed), or in some other way. Classical test theory has long provided psychometric definitions of parallelism (e.g., equal means, standard deviations, item intercorrelations, reliability coefficients, and factor structures across tests; Nunnally & Bernstein, 1994) as well as various models of parallelism, such as strictly parallel measures (equal true scores and variances), tau-equivalent measures (equal true scores and unequal variances), and essentially tau-equivalent or congeneric measures (linearly equatable true scores and unequal variances; Lord & Novick, 1968). Relatively speaking, it is easy to achieve parallelism when a measure covers a narrow content domain, such as a vocabulary test for a class or a knowledge test within a professional exam. However, for tests with more heterogeneous content, alternate forms must be carefully considered from both psychometric and substantive perspectives.

Situational Judgment Inventories

This article focuses on the latter type of test, specifically, the Situational Judgment Inventory (SJI), in which typically, each question reflects a hypothetical situation with multiple possible responses to that situation. Test takers are asked to indicate what they would most likely and least likely do in the given situation, and their responses are compared against the keyed responses of a reference group (e.g., experts who responded to the SJI). The content of the responses tends to be complex and varied—even within a single response—meaning that item responses may reflect a general factor of situational judgment yet contain unique situation-specific content (see Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004). The SJI of particular focus in this article was written to assess college students' decision making and judgment in 12 areas considered important to their success. Table 1 presents an example of the type of items included in this instrument. Individuals answering a, c, or e for "most likely" and b or f for "least likely" would agree with the experts' responses and would receive the highest score for this item (i.e., +2); conversely, individuals answering b or f for "most likely" and a, c, or e for "least likely" would disagree with the experts' responses and would receive the lowest score for this item (i.e., −2). On most items, individuals would receive scores somewhere between these two extremes.

Approaches for Generating Alternate Test Forms

One solution for writing alternate SJI items is "item cloning," taking each SJI test item and writing other items whose content and response options are paraphrased or otherwise very similar (Clause, Mullins, Nee, Pulakos, & Schmitt, 1998). Content similarity is obvious here, and it may not be possible to generate a large number of alternate forms. Although similar psychometric properties across alternate test forms are quite likely with item cloning, even meeting strict demands of psychometric parallelism, it may not be a satisfactory long-term approach for maintaining test-item security because knowledge of the item clones' content and nature would provide the answer to any of the individual clones.

More general procedures for generating alternate test forms are possible and may be especially useful when the test is relatively short yet the content domain represented by the sampled items is relatively complex (e.g., item content reflecting broad multidimensional constructs such as adaptability or general cognitive ability). Armstrong, Jones, and Wu (1992) applied network-flow algorithms that generate alternate forms

Table 1
Example of a Situational Judgment Item Measuring Leadership

An important class project you have been working on with a group of other students is not developing as it should because of petty differences and the need of some members to satisfy their own agenda. How would you proceed?

a. Try to solve the group problems before starting on the work. (+1)

b. Work hard by yourself to make sure the project is finished, taking on others' share of the work if necessary. (–1)

c. Talk to the professor and get suggestions about solving the problem. If that doesn't work, try to switch groups or have an independent project. (+1)

d. Schedule a number of meetings, forcing the group to interact. (0)

e. Take charge and delegate tasks to each person. Make them responsible for their part of the project. (+1)

f. Talk to the group and demand that they start working together. (−1)

What are you most likely to do?
What are you least likely to do?

Note. The numbers at the end of each response option have been added to the original item, indicating whether experts tended to agree that this was the most likely response (+1), the least likely response (–1), or neither (0).

by minimizing an overall discrepancy function based on a set of user-specified statistical criteria. Results suggest that their test-generating program based on these algorithms provided acceptable parallel forms (in terms of the mean, variance, reliability coefficient, and item response theory [IRT] test information function), though the method for constructing forms is somewhat indirect because their algorithm keeps permuting items until the discrepancy function is within a prescribed level of tolerance. In many cases, the required computer time is excessive even by today's standards (see Armstrong et al., 1992, pp. 279, 285).

Computer adaptive testing is another test development approach with the potential to create alternate test forms calibrated to each examinee while having similar psychometric characteristics. There are several reasons why computer adaptive testing may not be technically or practically feasible in some circumstances, however. First, computer adaptive testing procedures are based on IRT, which assumes unidimensionality of the item pool (McDonald, 1999). Although SJIs such as the one employed in this research rarely exhibit empirical evidence of multidimensionality, the percentage of variance accounted for by the first general factor in exploratory analyses tends to be low (often less than 20% of the total variance), and the content areas within an SJI often vary widely. Second, there are still many situations in which computers or computer adaptive tests are impractical and not accessible or affordable for large-scale administration. Third, at least when new tests are being developed, a calibrated item pool is often not large enough to allow for successfully implementing computer adaptive testing (we recognize that a similar, though less extreme situation, applies to the approach described in this article). Fourth, the complexities of IRT and computer adaptive testing still remain difficult to explain to a skeptical public that seems increasingly wary of the use of stan-

dardized testing (e.g., the recent controversy regarding the SAT generated by Atkinson, 2001). We believe the approach to the construction of alternate forms described here is relatively straightforward and would be practical in many applied contexts.

Extending the Gibson and Weiner Approach

The present illustration extends the basic approach offered by Gibson and Weiner (1998). This approach samples representative content from a diverse pool of items and generates a large set of preliminary test forms; user-defined criteria are then applied directly to select alternate forms from this set (where the user might typically be a test specialist or psychometrician). This method may be difficult to apply if the item pool is empirically multidimensional (Clause et al., 1998), but conversely, the method is often feasible if the test is empirically unidimensional. In this method, the psychometric properties of the items may differ, and there is likely more than one content area sampled as well. However, alternate forms, or even some form of psychometric parallelism, may be a reasonable assumption at the test level, the level at which scores are used operationally for making decisions (e.g., in personnel selection and educational settings; Raju, van der Linden, & Fleer, 1995).

Gibson and Weiner (1998) provided formulas and an example in which they generated parallel forms of a licensing exam, in which items within each parallel form were taken by different groups of examinees. Specifically, they combined items selected randomly from each content domain of the exam to create large numbers of preliminary forms. They then selected those forms whose classical test theory item statistics (i.e., mean, standard deviation, coefficient alpha) met their criteria for parallel forms. The present study employs the Gibson and Weiner approach in developing alternate forms for our SJI, using item statistics available from administrations of two separate versions of the test on two independent samples of college students. We also extend the Gibson and Weiner approach by incorporating data on the correlation between each item and the criterion of student grade point average (GPA). For each alternate test form, this allowed us to estimate not only the reliability but also the criterion-related validity of each test form. Finally, although our item pool is small, we demonstrate a method that investigates the adequacy of alternate forms with regard to high-stakes situations, in which security is a concern. Specifically, the method evaluates the extent to which our criteria for selecting alternate forms may have resulted in either (a) a set of forms that overused some sets of items more than others or (b) a set of forms containing different types of alternate tests (i.e., where each group of alternate forms tends to use different sets of items).

Method

Sample

This study used two samples of undergraduate freshmen at a large midwestern university who were recruited through their classes, housing units, and the student newspaper. In the spring semester of 2002, 654 freshmen volunteered for the study, receiving $40 for their participation. Of these, 644 provided usable data after screening for

careless responses on the first form of the SJI (containing 57 items). In the fall semester of 2003, another sample of 390 freshmen participated for extra credit in their introductory psychology course, with 381 providing usable data after screening for careless responses on the second form of the SJI (containing 96 items, none of which overlap with items in the first form). The two samples here differed in their external reward for participating (money vs. course credit), but both were college freshmen at the same university, both had similar GPAs as noted below, and both had similar demographics. Across samples, 76% of the participants were female, and the mean age was 18.4 years (*SD* = 0.62 years) with a positive skew: 66% of the sample was 18, and 32.6% of the sample was 19. The racial/ethnic composition was 79.7% Caucasians, 7.9% African American, 5.4% Asian, and 7% other.

Measures

SJI. In short, the SJI measure in this study resulted from a process of reviewing the literature on academic performance, categorizing the expressed goals of college mission statements, and interviewing residence life personnel. This process led to identifying 12 dimensions representing major criteria for college success within intellectual, interpersonal, and intrapersonal domains (see Table 2 and Oswald et al., 2004). The SJI we subsequently developed reflects hypothetical situations that correspond with these dimensions. First, items relevant to these dimensions were selected from the existing SJI literature and, when possible, were adapted to our purposes. Then, to supplement this, students provided written examples of the challenges and opportunities that 1st-year college students face with respect to these 12 dimensions. Item stems summarizing these responses were made, and an independent group of students was asked to describe how they might respond to each item stem. Next, students from an undergraduate psychological measurement course composed of juniors and seniors were asked to rate each response option on its effectiveness in the college context. These students were used as "expert" raters because the broad sample would hopefully reflect expertise across all 12 college success dimensions, not just the more traditionally academically oriented dimensions, and because their tenure reflected persistence and experience in the college setting.

Given these ratings, a scoring key was empirically developed to reflect the fact that individuals taking the SJI should get a higher score on items on which they tend to agree with student experts on the best and worst responses to a situation, as previously explained. Scores for each item range from –2 to +2 and reflect the SJI scoring procedure developed and reported in Motowidlo, Russell, Carter, and Dunnette (1988). Details of the scoring procedure are available from the first author.

The first phase of collecting data for the SJI involved the first sample mentioned above, for which we originally developed 155 items. Then, we omitted items based on (a) lack of rater agreement on the dimension to which the item should belong, (b) content redundancy with other items, and (c) conceptual clarity of item content or response options with respect to the dimension the item intends to measure. The goal was to have approximately 5 items per dimension, but for the dimension of career orientation, all 3 items were retained because there were too few to allow for any item selection. From the resulting 58 items, 3 did not have acceptable scoring keys on the basis of the expert ratings, 1 item was simply dropped, and replacements were obtained for the remaining 2 items, resulting in a final set of 57 SJI items.

Table 2
Twelve Dimensions of College Performance

Dimension	*Description*
Intellectual behaviors	
Knowledge, learning, and mastery of general principles (Knowledge)	Gaining knowledge and mastering facts, ideas, and theories and how they interrelate and understanding the relevant contexts in which knowledge is developed and applied. Grades or GPA can indicate, but not guarantee, success on this dimension.
Continuous learning and intellectual interest and curiosity (Learning)	Being intellectually curious and interested in continuous learning. Actively seeking new ideas and new skills, both in core areas of study as well as in peripheral or novel areas.
Artistic cultural appreciation and curiosity (Artistic)	Appreciating art and culture, either at an expert level or simply at the level of one who is interested.
Interpersonal behaviors	
Multicultural tolerance and appreciation (Multicultural)	Showing openness, tolerance, and interest in a diversity of individuals (e.g., by culture, ethnicity, or gender). Actively participating in, contributing to, and influencing a multicultural environment.
Leadership (Leadership)	Demonstrating skills in a group, such as motivating others, coordinating groups and tasks, serving as a representative for the group, or otherwise performing a managing role in a group.
Interpersonal skills (Interpersonal)	Communicating and dealing well with others, whether in informal social situations or more formal school-related situations. Being aware of the social dynamics of a situation and responding appropriately.
Social responsibility, citizenship, and involvement (Citizenship)	Being responsible to society and the community and demonstrating good citizenship. Being actively involved in the events in one's surrounding community, which can be at the neighborhood, town/city, state, national, or college/university level. Activities may include volunteer work for the community, attending city council meetings, and voting.

(continued)

A follow-up data collection was conducted one semester after the original data collection phase. The Situational Judgment Inventory Follow-Up (SJI-F) was developed after the 57-item SJI just described. Of the 98 items that went untested in the first SJI,

Table 2 (continued)

Dimension	*Description*
Intrapersonal behaviors	
Physical and psychological health (Health)	Possessing the physical and psychological health required to engage actively in a scholastic environment. This would include participating in healthy behaviors, such as eating properly, exercising regularly, and maintaining healthy personal and academic relations with others, as well as avoiding unhealthy behaviors, such as alcohol/drug abuse, unprotected sex, and ineffective or counterproductive coping behaviors.
Career orientation (Career)	Having a clear sense of career one aspires to enter into, which may happen before entry into college or at any time while in college. Establishing, prioritizing, and following a set of general and specific career-related goals.
Adaptability and life skills (Adaptability)	Adapting to a changing environment (at school or home), dealing well with gradual or sudden and expected or unexpected changes. Being effective in planning one's everyday activities and dealing with novel problems and challenges in life.
Perseverance (Perseverance)	Committing oneself to goals and priorities set, regardless of the difficulties that stand in the way. Goals range from long-term goals (e.g., graduating from college) to short-term goals (e.g., showing up for class every day even when the class isn't interesting).
Ethics and integrity (Ethics)	Having a well-developed set of values and behaving in ways consistent with those values. In everyday life, this probably means being honest, not cheating (on exams or in committed relationships), and having respect for others.

Note. The summary label for each dimension is in parentheses. These labels are used in Table 3.

96 items were categorized into dimensions and included in the SJI-F (2 items were dropped due to content irrelevance). The SJI-F was administered to the second sample previously described. For both SJI samples, exploratory and confirmatory factor analyses were conducted to test whether the items could be described empirically by the 12 dimensions to which they were initially assigned. Internal consistency reliabilities for the subscales for the SJI items for the first and second samples (SJI and SJI-F, respectively) were both low, with alphas ranging from .08 to .67. Exploratory factor analyses of the items supported a general factor (i.e., the first factor accounted for 3 times the

Table 3
Situational Judgment Inventory (SJI) and Situational Judgment Inventory Follow-Up (SJI-F): Scale Interitem Correlations and Composite Reliabilities

	SJI		*SJI-F*		*Total Item Bank*
Dimension	k	$\bar{r}_{ii'}$	k	$\bar{r}_{ii'}$	k
Knowledge	3	.073	3	.054	6
Learning	5	.026	2	.021	7
Artistic	5	.085	0	—	5
Multicultural	5	.048	12	.042	17
Leadership	5	.053	9	.030	14
Interpersonal	4	.053	16	.025	20
Citizenship	5	.035	5	.012	10
Health	5	.050	5	.033	10
Adaptability	5	.041	19	.019	24
Perseverance	5	.055	8	.044	13
Ethics	6	.063	10	.021	16
	k	α	k	α	k
Composite of items	57	.85	93	.88	150
Composite of scales	12	.83	12	.81	

Note. $\bar{r}_{ii'}$ = average interitem correlation for the dimension. For the SJI, $n = 640$ for the scales and composite of scales, and $n = 613$ for the composite of items. For the SJI-F, $n = 367$ for the scales and composite of scales, and $n = 314$ for the composite of items.

variance of subsequent factors), but similar to previous work with SJI measures, the percentage of variance accounted for by the first factor was less than 15% of the total for both the SJI and SJI-F sets of items. Table 3 presents the average interitem correlations of the items within each of the 12 dimensions used in constructing alternate forms. Coefficient alpha reliabilities at the composite levels were high: .85 and .88 at the item level and .83 and .81 at the scale level.

GPA. Students provided permission to request first-semester freshman GPA on a 4-point scale, which was provided by the registrar's office. For those in the first sample providing GPAs ($n = 619$), the mean GPA was 3.02 ($SD = 0.69$), and its criterion-related validity with the SJI item composite was $r = .15$ ($n = 617$ with both SJI and GPA, $p < .01$). For the second sample providing GPAs ($n = 368$), the mean GPA was 3.09 ($SD = 0.63$), and criterion-related validity with the SJI-F item composite was $r = .14$ ($N = 368$ with both SJI-F and GPA, $p < .01$).

Procedure

We followed Gibson and Weiner's (1998) multistep procedure for developing alternate forms of a test. Given the set of items from our 2 SJI measures, we carried out the procedure as follows. First, for all SJI items, a data file was created in which each row contained the following statistics for each item: the item mean, item standard deviation, and the item-total correlation uncorrected for the overlap between the item and total test score. Note that the item statistics are bound to the particular sample of items and individuals associated with a test administration. For instance, in our particular

example, some item statistics were based on the sample and item set associated with the first 57-item SJI form; other item statistics were based on the sample and item set associated with the second 96-item SJI-F form. These item statistics can be subsequently refined and updated in the data file as items appear on subsequent test forms and are administered to new samples.

Second, we extended the Gibson and Weiner (1998) procedure by adding to the data file the correlation between each item and a criterion (in this case, SJI items and student 1st-year GPA, respectively), which allowed for criterion-related validity estimates for each alternate test form. This may seem like a minimal extension of the procedure, but it is quite a critical one because it leads to test forms that are sensitive not only to the internal consistency of the measure but also to the extent that the measure will relate to measures of other constructs. Both the internal and external concerns are important when trying to understand the theoretical relevance of a construct as reflected in a measure (Cronbach & Meehl, 1955). The third step required grouping the rows (items) in the data file into each of the 12 dimensions (e.g., all items for the leadership dimension were listed together). Although the SJI data for both tests empirically supported unidimensionality, alternate SJI test forms should still contain representative content by sampling the same number of items within each dimension.

The next step in generating alternate forms required producing a large set of preliminary test forms from which to select the set of acceptable alternate forms. To that end, an SAS/IML program (SAS Institute, Inc., 1999) was written that created 10,000 preliminary SJI test forms. Each preliminary test form contained 36 items, resulting from taking the total item pool (combining the SJI and SJI-F items) and randomly sampling 3 items within each of the 12-item sets for each dimension of college performance. Sampling 3 items per dimension helped ensure adequate content sampling, and based on the item-composite results from Table 3, we estimated that coefficient alpha reliability would be adequate, averaging about .76. For each preliminary test form, the program estimated the mean, standard deviation, coefficient alpha, and criterion-related validity, which can be derived from the item-level statistics as follows.

Given a preliminary test form X, the mean of the test scores was estimated as the sum of the means for each of its k items $t_1, t_2, \ldots t_k$:

$$\bar{X} = \sum_{i=1}^{k} \bar{t}_i. \tag{1}$$

Note that each individual item t_i can come from different tests T or from different test administrations. Gibson and Weiner (1998) found that the test to which an item belongs has a small and random effect on item statistics. Thus, the database of item statistics can be continually refined as an item is used across different alternate test forms and test administrations.

The standard deviation of the test scores is estimated by summing the product of each item-total correlation with its respective item standard deviation:

$$s_x = \sum_{i=1}^{k} r_{t_i T_i} s_{t_i}. \tag{2}$$

This is not the corrected item-total correlation, in which the total score subtracts out the item score. All item-total correlations must include the item score in the total score so that item variances (not just covariances) contribute toward the estimate of the variance (or standard deviation) of the test.

Next, coefficient alpha for the test is estimated in the usual way, based on the total test variance and the sum of the item variances:

$$\alpha = \left(\frac{k}{k-1}\right)\left(\frac{s_x^2 - \sum_{i=1}^{k} s_{t_i}^2}{s_x^2}\right). \tag{3}$$

Finally, the criterion-related validity (or validities if there are several criteria) of the preliminary test form was estimated from a formula provided by Ghiselli, Campbell, and Zedeck (1981, p. 433; see also Humphreys, 1956):

$$r_{XY} = \frac{\sum_{i=1}^{k} r_{t_i Y} s_{t_i}}{\sum_{i=1}^{k} r_{t_i T_i} s_{t_i}}. \tag{4}$$

These statistics were estimated for 10,000 preliminary test forms and are summarized in Table 4.

Results and Discussion

Given these statistics for the preliminary test forms, we then applied criteria for identifying desirable alternate forms (see Table 5). Many of these criteria are consistent with psychometric standards for parallel tests: First, we wanted the test means to be similar, so one criterion was for the standardized effect size between the preliminary form and the overall mean across forms to be within $|d| \leq .05$. This criterion alone eliminated most (about 96%) of the preliminary test forms, though obviously a less stringent criterion would result in including more tests (e.g., $|d| = .10$ might be acceptable given that .20 is defined as small by Cohen, 1988, p. 25). Second, we wanted the alpha reliabilities of the alternate forms to meet the conventional standard of being at or above .70. Third, extending the Gibson and Weiner (1998) approach, we applied a criterion-related validity standard: To be considered an alternate test form, each preliminary test form needed to have a criterion-related validity with GPA of $r_{xy} \geq .15$. This value is somewhat arbitrary but is the mean validity across preliminary test forms and could be viewed as a reasonable practical minimum criterion-related validity for a test. Of the tests remaining, a final criterion was to trim the outlying 20% of standard deviation values out of the distribution (i.e., trimming 10% off both tails of the distribution of preliminary forms). This left us with a set of 144 tests that, as defined by these selection criteria, are alternate forms. Note that although our example results in

Table 4
Descriptive Statistics for Preliminary Test Forms and Selected Parallel Test Forms

Test Statistic	M	SD	*Minimum*	*Maximum*
Characteristics of 10,000 preliminary test forms				
$\bar{X}$	26.07	2.08	18.05	33.82
s_x	10.58	0.75	7.85	13.82
α	.72	.03	.52	.82
r_{xy}	.15	.03	.03	.26
Characteristics of 144 selected parallel test forms				
$\bar{X}$	26.06	0.07	25.96	26.17
s_x	10.80	0.41	10.10	11.56
α	.73	.02	.70	.78
r_{xy}	.17	.02	.15	.23

Note. For the Situational Judgment Inventory, $n = 640$ for the scales and composite of scales, and $n = 613$ for the composite of items. For the Situational Judgment Inventory Follow-Up, $n = 367$ for the scales and composite of scales, and $n = 314$ for the composite of items.

Table 5
Criteria for Selecting Parallel Forms From the Preliminary Forms

Selection Criterion	*Number of Preliminary Test Forms Remaining*
No selection criteria	10,000
Test means within $\lvert d \rvert \leq .05$ of the overall mean	421
$\alpha \geq .70$	339
$r_{xy} \geq .15$	173
Test *SD*s are within the central 80% of the distribution of the *SD*s of the preliminary test forms (i.e., trim outer 20% of the *SD*s)	144

144 alternate forms, we could have generated more alternate forms by first generating a larger set of preliminary forms. Based on our item bank, there were 5.16×10^{26} possible unique preliminary forms—much more than the 10,000 preliminary forms we generated.

Given the final set of 144 alternate forms, we wanted to investigate the similarity of these forms further. Ideally, alternate forms are relatively interchangeable with one another, and items would be equally likely to appear on any form. Hypothetically, we could find that the 144 alternate forms met the criteria we set out, but (a) different groups of test forms share different sets of items or (b) different groups of items appear on different forms of the test. Evaluation of the first concern was assessed by generating a 144×144 matrix of the overlap between all pairs of alternate forms, in which each cell in the matrix reflected the percentage of items shared between a pair of forms, out of the total number of items. Referring to Table 6, the mean item overlap was 30%, which may be high given test-security concerns but could be considered relatively low

Table 6
Further Evidence for Parallel Forms

	Eigenvalue (% of Variance)	
Analysis	*First Factor*	*Second Factor*
Factoring % test-pair similarity across items, mean similarity = 30%, *SD* = 6%, min = 8%, max = 53%	43.0 31.0	2.9 2.1
Factoring % item-pair agreement across tests, mean agreement = 5.3%, *SD* = 5%, min = 0%, max = 46%	11.2 7.5	1.4 1.0
Item pair agreement: observed % – chance agreement %; mean difference = –0.2%, *SD* = 2.94%, min = –16.6%, max = 19.4%		

Note. A total of 150 items, 144 tests. Principal axis factor analysis without rotation. Factoring test-pair similarities was conducted on the first 139 of the 144 tests in the database, as adding more tests made the matrix nonpositive definite. Similar results came from factoring subsets of tests that included these last 5 tests (e.g., the last 139 tests in the database).

given our small item bank. In addition to this analysis, principal axis factor analysis was applied to the item-pair agreement matrix, and the unrotated solution indicated a large general factor (i.e., the first eigenvalue was much larger than the second eigenvalue), providing further support that the alternate forms are similar, with no systematic subgroups of test forms sharing different sets of items with one another.

The second concern regarding item clustering was addressed in two ways. The first way was to generate a 150×150 matrix of the pairing of items that occurred across all alternate forms, where pairing reflected the extent to which a pair of items appeared across all forms. Table 6 shows that mean item pairing was 5.3%, and therefore, most items did not appear consistently with any other particular item. Although there was positive skew for number of item pairings (the maximum was 46%), a principal axis factor analysis applied to this percentage-pairing matrix yielded a large general factor, as in the previous analysis. In this case, a large general factor indicates that the process of selecting alternate forms did not lead to items appearing on tests systematically with some items but not others. The second way we examined this is related to the first way but perhaps is a bit more specific: We took into account the fact that some items would be more likely to appear together if they came from the same dimension, especially for those dimensions in which there are very few items. Therefore, we calculated the observed percentage pairing between items and subtracted from it the percentage pairing that would be expected due to chance. Results in Table 6 suggest that most items appeared as often as what one might expect due to chance, with the mean difference being only –0.2%. The minimum and maximum values in the distribution suggested that some items co-occurred more (or less) often than what pure chance would have predicted, and this should be investigated further, but generally, this analysis combined with the previous factor analysis suggested that items do not tend to occur more often with one another as a result of the process of selecting the set of alternate forms from the total set of preliminary test forms that best meets our criteria.

However, if we were to continue with this method, actually implementing an SJI with multiple alternate forms with high test security, we would certainly not stop at this point. Large item banks and independent test forms help prevent cheating as well as minimize the impact of coaching individuals on specific item content. We would still need to develop more SJI items for all dimensions and once again apply the method of factor analysis to detect whether systematic types of forms or item clusters are found in the new alternate forms. New items would be developed and rated on their content relevance, as was done in the present study, and then once an item was administered in a new SJI measure, the Gibson and Weiner (1998) method would be used to examine the influence of new item statistics on the test statistics of new candidate preliminary forms to be considered for selection as alternate forms.

That said, the results of the present analyses leave us satisfied that the 144 alternate test forms selected are relatively parallel to one another psychometrically, contain distinct sets of items, and show adequate criterion-related validity for predicting GPA. We developed alternate test forms by considering test-level statistics, not item statistics, with the former being the level at which important decisions are made (e.g., personnel selection, college admissions). Consequently, whole test forms must be generated and psychometrically screened, but the computer power required to do that is more than adequate (e.g., in the present study, test form generation and subsequent estimates of test-level statistics were available almost instantaneously). As previously mentioned, item statistics can be refined over time, and the more an item is administered within more test forms to more diverse samples, the test statistics estimated from such items become, depending on one's purpose, either more robust (one can aggregate across samples) or more specific (one can have specific information for specific types of samples). Clearly, if the samples associated with an item were small, then there would be legitimate concern for finding larger samples or for incorporating some sort of a cross-validation method into the procedure for generating alternate forms. Regarding this issue for the present study, given similar test data—say, an effective sample size of 400 and a test length of 36 items—a value of $\alpha = .70$ would yield a 95% confidence interval ranging from .66 to .75. Computation of this interval assumes that tests are essentially parallel (Hakstian & Whalen, 1976), but the Type I error associated with such a confidence interval is very close to nominal levels even when items are not fixed, so long as the number of items is about 40 (Hakstian & Barchard, 2000).

We selected alternate test forms largely based on traditional psychometric characteristics, but we extended this to include a standard for criterion-related validity. In fact, one could develop an item database to include additional characteristics on which to generate candidate test forms, then develop and apply different selection criteria to obtain alternate forms. Such criteria could be based on

- test score means that are near the cutoff point at which one is making decisions (e.g., in employment, college admissions, or licensure contexts);
- similar distributions of item difficulties as suggested by classical test theory or IRT (see Armstrong, Jones, & Wang, 1994; van der Linden, 1998);
- race differences, gender differences, or other mean differences on the test (estimated from item-mean differences and item intercorrelations; relevant formulas can be found in Rosenthal & Rubin, 1986; Sackett & Ellingson, 1997);
- expert judgments about the social desirability of the item or correlations with a social desirability measure;

- test takers' reactions to the test, based on reactions (or ratings of anticipated reactions based on invasiveness, fairness, etc.) across constituent items; or
- criterion-related validity coefficients across multiple criteria, not just one criterion.

van der Linden (1998) reviewed other criteria and methodological approaches to test-form development, whereas Wightman (1998) outlined other criteria in addition to many practical constraints and costs in test-form development. From the Gibson and Weiner (1998) approach we adopted and extended, we are careful to say that the procedure generates alternate forms, not strictly parallel forms in the psychometric sense. In this procedure, items can be given to entirely different samples and be in entirely different test forms, which makes explicit the fact that item responses are a product of item-sample interactions (among other things). These interactions introduce random effects into reliability and validity estimates, but if you believe that test forms will in fact be administered in different forms to different samples, this potentially leads to more robust estimates of overall reliability and validity. Corroborating this approach to test development, Stanton, Sinar, Balzer, and Smith (2002) noted that psychometric characteristics internal to the test should not be the only important concern (or hurdle) when developing test forms. Simply lengthening test forms may improve alpha coefficients yet increase administration time and cost as well as lead to test-taker fatigue. Furthermore, a good bank of test items requires writing items based on a solid definition and understanding of the content domains of interest. Creating items within a scale that are redundant with one another may increase alpha coefficients yet lead to deficient sampling of the intended content domain as well as decreased criterion-related validity. The present example sought high estimates of alpha reliability for SJI items that were complex in nature. Items loaded on a general factor, yet there was evidence for complexity within each item (i.e., low item factor loadings) and the content of the SJI items was rationally developed across 12 different dimensions of college performance. In other words, development of alternate forms balanced the concerns for high reliability along with representative sampling of item content across the 12 dimensions. Neither the alpha coefficients nor the evidence for a weak general factor imply unidimensionality (Cortina, 1984), although they are sufficient conditions for the procedure applied here.

In addition to the internal characteristics of a test, McDonald (1999) noted that rarely are parallel forms examined for similar correlations with meaningful external variables. The present study conducted this sort of examination by extending the Gibson and Weiner (1989) procedure to incorporate criterion-related validity, specifically, item-criterion correlations between the SJI and GPA. There are several reasons to use all SJI items and develop SJI composite forms rather than just the Knowledge and Continuous Learning SJI items (Tables 2 and 3) to develop an intellectual scale that relates to GPA. First, empirical support lent itself only to the general factor of situational judgment, and it is quite sensible to think that possessing good judgment across a variety of situations (intellectual in nature or otherwise) would correlate with GPA. Second, the small item bank precluded a scale-level analysis. Doing so would have led to low alpha reliability coefficients and/or not enough items to demonstrate the method of generating parallel forms, which was the main thrust of the article. Third, although GPA is based largely on knowledge, it is also clearly effort based. Keeping up good grades over time can in part be attributed to motivational and noncognitive factors included in the SJI composite: proximal factors such as perseverance (going to

class on time, turning in assignments even when dealing with serious personal matters) or more distal factors such as physical and psychological health (not having a nervous breakdown, not having to go to the hospital), interpersonal skills (working effectively in a lab group), and ethics and integrity (not plagiarizing, not cheating on an exam).

Extending this approach to include some of the additional criteria mentioned above would also involve a consideration of how a test functions in its relationships with external variables (e.g., test taker reactions, social desirability), and although SJI items loaded on a general factor, item composites based on the 12 SJI dimensions may well demonstrate interpretable patterns of differential validity with respect to criteria. In short, the validation of a measure—and alternate or parallel forms of a measure—is a continuous process requiring the accumulation and evaluation of evidence from a variety of sources (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The Gibson and Weiner (1998) approach, together with our extension to criterion-related validity offered in this article, provides a straightforward, practical, and psychometrically sound procedure for researchers and practitioners seeking to develop and validate alternate forms of a measure containing diverse content areas.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Armstrong, R. D., Jones, D. H., & Wang, Z. (1994). Automated parallel test construction using classical test theory. *Journal of Educational Statistics*, *19*, 73-90.

Armstrong, R. D., Jones, D. H., & Wu, I.-L. (1992). An automated test development of parallel tests from a seed test. *Psychometrika*, *57*, 271-288.

Atkinson, R. C. (2001, April). *Standardized tests and access to American universities*. The 2001 Robert H. Atwell Distinguished Lecture, delivered at the 83rd Annual Meeting of the American Council on Education, Washington, DC.

Clause, C. S., Mullins, M. E., Nee, M. T., Pulakos, E., & Schmitt, N. (1998). Parallel test form development: A procedure for alternate predictors and an example. *Personnel Psychology*, *51*, 193-298.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology*, *78*, 98-104.

Cronbach, L. J., & Meehl, P. C. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: Freeman.

Gibson, W. M., & Weiner, J. A. (1998). Generating random parallel test forms using CTT in a computer-based environment. *Journal of Educational Measurement*, *35*, 297-310.

Hakstian, A. R., & Barchard, K. A. (2000). Toward more robust inferential procedures for coefficient alpha under sampling of both subjects and conditions. *Multivariate Behavioral Research*, *35*, 427-456.

Hakstian, A. R., & Whalen, T. E. (1976). A *k*-sample significance test for independent alpha coefficients. *Psychometrika*, *41*, 219-231.

Humphreys, L. G. (1956). The normal curve and the attenuation paradox in test theory. *Psychological Bulletin*, *53*, 472-476.

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

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Motowidlo, S. J., Russell, T. L., Carter, G. W., & Dunnette, M. D. (1989). *Revision of the management selection interview: Final report*. Minneapolis, MN: Personnel Decisions Research Institute.

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

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, *89*, 187-207.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*, 353-368.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, *99*, 400-406.

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*, 707-721.

SAS Institute, Inc. (1999). *SAS/IML user's guide*. Version 8. Cary, NC: Author.

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*, 167-194.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*, 195-211.

Wightman, L. F. (1998). Practical issues in computerized test assembly. *Applied Psychological Measurement*, *22*, 292-302.

Frederick L. Oswald obtained his Ph.D. from the University of Minnesota in 1999 in industrial/organizational psychology and is currently an assistant professor of psychology at Michigan State University. His research deals with person-job fit as it relates to employment and academic settings, and in collaboration with others at Michigan State University and the College Board, he is conducting work on noncognitive measures in the college admissions context. His methodological work includes evaluations and extensions of meta-analytic approaches.

Alyssa J. Friede is a graduate student in industrial/organizational psychology at Michigan State University, having graduated summa cum laude from the University of Pennsylvania in 2002. Her research focuses on noncognitive predictors of college student success and balance between work and nonwork roles.

Neal Schmitt obtained his Ph.D. from Purdue University in 1972 in industrial/organizational psychology and is currently University Distinguished Professor of Psychology and Management at Michigan State University. He was editor of the Journal of Applied Psychology *from 1988 to 1994 and was president of the Society for Industrial and Organizational Psychology from 1989 to 1990. He has received the Society for Industrial/Organizational Psychology's Distinguished Scientific Contributions Award (1999) and its Distinguished Service Contributions Award (1998). He was also awarded the Heneman Career Achievement Award from the Human Resources Division of the Academy of Management. He has coauthored three textbooks and published approximately 150 articles. Over the past 3 years, he has also been working on the development and validation of noncognitive measures for college admissions.*

Brian H. Kim received his M.A. in industrial/organizational psychology from Michigan State University in 2004 and is currently working toward his Ph.D. His current work involves research on noncognitive measures for college admissions, adaptive performance in work teams, and response surface methodologies as they apply to person-job fit research.

Lauren J. Ramsay obtained her M.A. in industrial/organizational psychology from Michigan State University in 2003 and is currently working toward her Ph.D. She has received jointly written student research grants from the College Board as well as a fellowship from the U.S. Department of Education for foreign language and area studies in Africa.