CollegeBoard

The Effect of Using Different Weights for Multiple-Choice and Free-Response Item Sections

Amy Hendrickson
Brian Patterson
Gerald Melican

The College Board

National Council for Measurement in Education
March 27th, 2008 New York

Combining Scores to Form Composites

- Many high-stakes tests include both multiple-choice (MC) and free response (FR) item types
- Linearly combining the item type scores is equivalent to deciding how much each component will contribute to the composite.
- Use of different weights potentially impacts the reliability and/or validity of the scores.

Advanced Placement Program® (AP®) Exams

- 34 of 37 Advanced Placement Program® (AP®) Exams contain both MC and FR items.
- Weights for AP composite scores are set by a test development committee.
- These weights range from 0.33 to 0.60 for the FR section.
  - These are translated into absolute weights which are the multiplicative constants that are applied directly to the item scores
- Previous research exists concerning the effect of different weighting schemes on the reliability of AP exam scores but not on the validity of the scores.

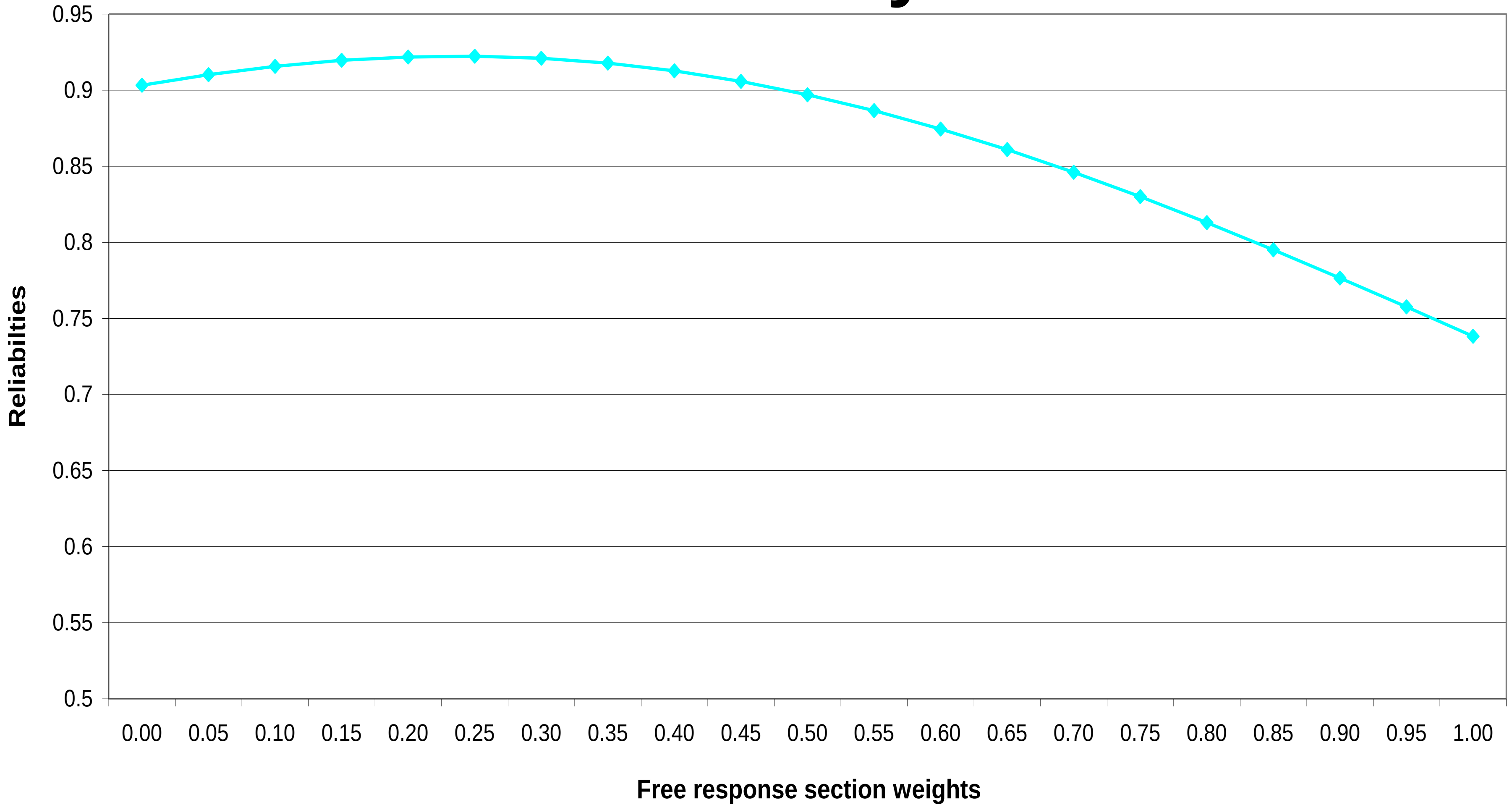
Weighting and Reliability and Validity

- Construct/Face validity perspective
  - Include both MC and FR items because they measure different and important constructs.
  - De-emphasizing the FR section may disregard the intention of the test developers and policy makers who believe that, if anything, the FR section should have the higher weight to meet *face* validity.
- Psychometric perspective
  - FR are generally less reliable than MC, thus, it is best to give proportionally higher weight to the (more reliable) MC section.
- 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).

Weighting Schemes

- Relative weights
  - based on desired proportion contribution of each item section to the total composite raw score.
  - nominal, logical, or *a priori* weights
- Item-score weights (Moses, Dorans, Miao, and Yon (2006))
  - applying the same weight to each FR item as is applied to each MC item
  - minimizes the influence of FR score unreliability, among other psychometric benefits
- Weights that maximize test score reliability (Gulliksen, 1950; Kolen, 2007; Walker, 2005)
  - empirically determine the set of weights that maximize the composite score reliability

Determining Weights for Maximum Reliability



Purpose of this Study

Compare the effect of the different weighting schemes (i.e., current, item-score and those that maximize total test reliability) on effective weights, test reliability, and validity coefficients for a selection of AP Exams.

Advanced Placement Exams

- Five AP Exams studied here
  - Macroeconomics
  - Environmental Science
  - Spanish Language
  - English Language
  - Spanish Literature

2006 AP Operational Exam and Examinee Specifications

	Macro-economics	Environmental Science	Spanish Language	English Literature	Spanish Literature
Relative Weights {MC, FR}	{0.667, 0.333}	{0.600, 0.400}	{0.500 , 0.500}	{0.450, 0.550}	{0.400, 0.600}
Total Points	90	150	180	150	150
Composite Score Mean	43.130	65.840	86.260	79.210	79.700
Composite Score SD	19.020	26.190	30.910	23.020	22.400
Composite Score SEM	5.410	6.960	9.290	8.210	8.520
Correlation (MC, FR)	0.765	0.787	0.781	0.622	0.672
Composite Score Reliability	0.919	0.929	0.910	0.873	0.855
FR Reliability	0.739	0.761	0.776	0.730	0.709
MC Reliability	0.903	0.925	0.897	0.862	0.851
Examinee Population	48,923	42,991	42,279	270,555	14,287
Examinee Sample	3,515	1,420	2,561	16,986	303

Methods – Reliability Analyses

- Applied three sets of weights (current, item-score, and maximum reliability) to the summary statistics of the operational test scores for each of the five AP Examinations.

- Computed:
  - Effective weights
    - Proportion of composite score variance that is attributable to a component of the composite (Kolen, 2007))
  - Coefficient alpha reliability values
    - Conducted the Feldt (1980) test of the difference between two reliability coefficients that are based on the same sample of examinees

Other Data

- SAT Reasoning Test scores (introduced in March of 2005)
- First-year college or university grade-point-average (GPA)
- Representative sample of 55 colleges and universities for all first-time, first-year students who started in the Fall of 2006
- Matched with AP scores

Methods – Validity Analyses

- Calculated AP composite test scores with each of the three sets of weights for each AP examinee.
- Correlated each AP score with:
  - SAT-Math (SAT-M)
  - SAT-Critical Reading (SAT-CR), and
  - SAT-Writing (SAT-W)
  - College first-year GPA
  - Correlations computed separately for each institution and a weighted average (using AP-exam-taking sample sizes for each institution) was computed.
  - Hotelling-Williams tests (Bobko, 2001)
    - Two-sided critical t-values based on n-3 degrees of freedom and a Bonferroni-adjusted $\tilde{\alpha}$ value were compared to each of the three tests of the differences between the correlations.

Relative and Effective Weights Results

	Operational Relative Weight		Effective Relative Weight	
Weighting Scheme	FR	MC	FR	MC
Macroeconomics				
Current	0.333	0.667	0.320	0.680
Item-Score	0.310	0.690	0.290	0.710
Maximum Reliability	0.250	0.750	0.230	0.770
Environmental Science				
Current	0.400	0.600	0.370	0.630
Item-Score	0.290	0.710	0.250	0.750
Maximum Reliability	0.250	0.750	0.220	0.780
Spanish Language				
Current	0.500	0.500	0.450	0.550
Item-Score	0.440	0.560	0.380	0.620
Maximum Reliability	0.300	0.700	0.240	0.760
English Literature				
Current	0.550	0.450	0.510	0.490
Item-Score	0.330	0.670	0.260	0.740
Maximum Reliability	0.350	0.650	0.280	0.720
Spanish Literature				
Current	0.600	0.400	0.580	0.420
Item-Score	0.400	0.600	0.340	0.660
Maximum Reliability	0.350	0.650	0.290	0.710

Differences in Reliability Values

	Macro- economics	Environmental Science	Spanish Language	English Literature	Spanish Literature
Summary Statistics					
Relative Weights {MC, FR}	{0.667, 0.333}	{0.600, 0.400}	{0.500, 0.500}	{0.450, 0.550}	{0.400, 0.600}
Correlation (MC, FR)	0.765	0.787	0.781	0.622	0.672
Multiple Choice Reliability	0.903	0.925	0.897	0.862	0.851
Number Multiple Choice Items	59	100	75	55	64
Sample size	3,515	1,420	2,561	16,986	303
Reliability Values					
Current Composite Reliability	0.919	0.929	0.910	0.873	0.856
Item-score Composite Reliability	0.921	0.937	0.916	0.891	0.883
Maximum Composite Reliability	0.922	0.938	0.920	0.891	0.885
Reliability Differences					
I-C Reliability Difference	0.002	0.008	0.006	0.018	0.027
M-C Reliability Difference	0.003	0.009	0.010	0.018	0.029
M-I Reliability Difference	0.001	0.001	0.004	0.000	0.002
Test Length Factor					
Current Composite Reliability	1.219	1.061	1.161	1.100	1.041
Item-score Composite Reliability	1.252	1.206	1.252	1.309	1.321
Maximum Composite Reliability	1.270	1.227	1.321	1.309	1.347
Additional MC Items					
Current Composite Reliability	13	6	12	6	3
Item-score Composite Reliability	15	21	19	17	21
Maximum Composite Reliability	16	23	24	17	22

Concurrent Validity Results

	Relative		Correlation of the Weighted Composite with		
Weighting Scheme	FR Weight	Reliability	SAT-M	SAT-CR	SAT-W
Macroeconomics (N=3,515)					
Current	0.333	0.919	0.448 [b,c]	0.415 [b,c]	0.340 [b,c]
Item-Score	0.310	0.921	0.451 [a,c]	0.419 [a,c]	0.343 [a,c]
Maximum Reliability	0.250	0.922	0.456 [a,b]	0.428 [a,b]	0.348 [a,b]
Environmental Science (N=1,420)					
Current	0.400	0.929	0.423 [b,c]	0.550 [b,c]	0.405
Item-Score	0.290	0.937	0.429 [a]	0.558 [a,c]	0.409
Maximum Reliability	0.250	0.938	0.430 [a]	0.560 [a,b]	0.409
Spanish Language (N=2,561)					
Current	0.500	0.910	0.143 [b,c]	0.243 [b,c]	0.258 [b,c]
Item-Score	0.440	0.916	0.148 [a,c]	0.258 [a,c]	0.265 [a,c]
Maximum Reliability	0.300	0.920	0.158 [a,b]	0.285 [a,b]	0.277 [a,b]
English Literature (N=16,986)					
Current	0.550	0.873	0.371 [b,c]	0.660 [b,c]	0.617 [b,c]
Item-Score	0.330	0.891	0.403 [a,c]	0.706 [a,c]	0.634 [a]
Maximum Reliability	0.350	0.891	0.402 [a,b]	0.704 [a,b]	0.634 [a]
Spanish Literature (N=303)					
Current	0.600	0.856	0.124	0.223	0.216 [b,c]
Item-Score	0.400	0.883	0.133	0.238	0.243 [a]
Maximum Reliability	0.350	0.885	0.134	0.241	0.248 [a]

Predictive Validity Results

	Relative		Correlation with
Weighting Scheme	FR Weight	Reliability	First-Year GPA
Macroeconomics (N=3,515)			
Current	0.333	0.919	0.386 [b,c]
Item-Score	0.310	0.921	0.387 [a,c]
Maximum Reliability	0.250	0.922	0.390 [a,b]
Environmental Science (N=1,420)			
Current	0.400	0.929	0.278
Item-Score	0.290	0.937	0.275
Maximum Reliability	0.250	0.938	0.273
Spanish Language (N=2,561)			
Current	0.500	0.910	0.200
Item-Score	0.440	0.916	0.202
Maximum Reliability	0.300	0.920	0.204
English Literature (N=16,986)			
Current	0.550	0.873	0.373 [b,c]
Item-Score	0.330	0.891	0.363 [a,c]
Maximum Reliability	0.350	0.891	0.365 [a,b]
Spanish Literature (N=303)			
Current	0.600	0.856	0.268
Item-Score	0.400	0.883	0.279
Maximum Reliability	0.350	0.885	0.281

Reliability Summary

- Use of different weighting schemes does impact the reliability of the scores for the AP Exams, especially for those with high FR section weights.
  - The increase in reliability by using the maximum reliability weighting scheme was as much as 0.029
  - May seem small in terms of the reliability value, translated into the number of MC items that would need to be added to cause this increase, the impact is more readily apparent.
- Consistent with previous findings

Validity Summary

- Inconclusive, but seems to indicate that use of different weighting schemes does impact the correlation of the scores from the AP Exams with both concurrent and predictive criteria, to a small extent.
  - Both the item-score and the maximum reliability schemes led to higher concurrent validity coefficients for 4 out of the 5 exams and higher predictive validity coefficients for 1 exam.
  - Current weighting scheme led to higher predictive validity for 1 exam.
  - The maximum reliability weighting scheme generally led to higher predictive and concurrent validity coefficients than the item-score weighting scheme, indicating both a reliability and validity advantage for this scheme.
- Increased reliability does not always lead to increased validity (Penev and Raykov, 2006).

Limitation

- Most appropriate concurrent and predictive criteria?
  - Correlation between disparate AP and SAT Exams (AP Spanish Literature and SAT–M).
  - First-year college GPA based on all student coursework in the first year. Courses vary in difficulty and in the extent to which the content matches the AP Exams to which we are correlating.

Future Work

- Examine the relationship between AP Exam scores with first-year college course-level GPA for matched subject areas (i.e., Spanish with Spanish).
- Examine the effect of different weighting schemes on the classification of students into the 1-5 scaled-score scale Modu (1981).
- Other weighting schemes:
  - Unweighted
  - IRT pattern weighting

Conclusions

The results are informative for large-scale mixed-format assessments in that they not only evaluate the differences in reliability resulting from different weights, but also on validity coefficients.

This addition to the relevant literature will provide new evidence to test developers and psychometricians alike as they develop new examinations and revise existing ones.

Questions, Comments, Suggestions

- Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board presentations do not necessarily represent official College Board position or policy.
- Please forward any questions, comments, and suggestions to:

  - ahendrickson@collegeboard.org,
  - bpatterson@collegeboard.org,
  - gmelican@collegeboard.org