Modeling the Predictive Validity of SAT Mathematics Items Using Item Characteristics

Jennifer L. Kobrin & Rachel Kim
The College Board

Paul Sackett
University of Minnesota

Collecting Validity Evidence at the Item Level

- Most studies on the validity of high-stakes tests examine the relationship between total test scores and various outcomes of interest.
- Yet, since the test item is the basic unit of observation in a test, it is important to collect validity evidence for both item responses and test scores (Haladyna, 2004).

Research Questions

(1) How do SAT mathematics items' content area, format, cognitive complexity, and abstract/ concrete designation affect the item's prediction of college outcomes?

(2) Is the relationship between item characteristics and college outcomes mediated by item difficulty and discrimination?

A Little About the SAT Math Test

- Consists of 54 items, administered in three separately timed sections.
- 44 multiple-choice and 10 student-produced response (SPR).
- Item content:

  20-25% Numbers & Operations (NO)

  35-40% Algebra &Functions (AF)

  25-30% Geometry & Measurement (GM)

  10-15% Data Analysis, Statistics, Probability (DA)

Additional Item Coding

- Cognitive Complexity – 3 Levels:
  - Routine (RTNE)
  - Comprehension (COMP)
  - Non-routine/insightful (NONR)
- Abstract (ABS) or Concrete (CONC)

Example Item #1

GM, SPR, RTNE, CONC

The floor of a rectangular room measures 16 feet by 12 feet. The floor is to be carpeted using square carpet tiles that measure 2 feet along each side. How many tiles are needed to completely cover the floor?

Example Item #2

GM, MC, Comp, ABS

If two of the three angles of an isosceles triangle have measures of 50° and 80°, respectively, what is the measure of the third angle?

(A) 50°

(B) 60°

(C) 65°

(D) 80°

(E) It cannot be determined from the information given.

Example Item #3

AF, MC, NONR, CONC

$$N(t) = 50 * 2^{t/12}$$

The function above gives the number of cells, *N*(*t*), in a certain culture *t* minutes after an experiment began. The number of cells in the culture 1 hour after the experiment began is how many times the number of cells at the beginning of the experiment?

(A) 16

(B) 32

(C) 60

(D) 800

(E) 1,600

Data Source

- 110 colleges and universities provided FYGPA, course grades, and major for 161,584 students who finished their first year of college in spring 2007.
- Official SAT scores obtained from 2006 College-Bound Senior cohort database.
- 644 items from 12 SAT forms administered between March 2005 and April 2006.

Outcome Variables

1. Point-biserial correlation of item score (1,0) with FYGPA **(FYGPA validity coefficient).**
2. Point-biserial correlation of item score (1,0) with average mathematics course grades **(Math course grade validity coefficient).**
3. Percentage of students majoring in a STEM field among those answering the item correctly **(STEM percentage)**.

Predictor Variables

1. A set of effect code variables to designate the 12 separate cohorts of students taking each SAT form.
2. Effect code variables for the items' content area, format, cognitive category, and abstract/concrete.
3. Measures of item difficulty and item discrimination.

Statistical Analyses

- Regression of FYGPA validity coefficients, math course grade validity coefficients, and STEM percentage on item characteristics
    - First set entered only item characteristics
    - Second set controlled for item difficulty and discrimination.
- In all models, cohort effect codes were entered first.
- Final models entered 2- and 3-way interaction effects (retained only if significant at $p < .01$)

RESULTS

Change in R-Square Associated with Each Predictor: FYGPA Model

Model	Predictors	No covariates	With covariates
1	Cohort	**.139***	**.139***
2	Add Item difficulty/ discrimination	--	**.443***
3	Add Abstract	**.057***	**.023***
4	Add Cognitive	**.041***	**.019***
5	Add Content	**.023***	.006
6	Add Format	**.016***	**.008***
7	Add 2-way interactions	.014	**.020***
8	Add 3-way interactions	.017	.015
Total R-Sq. **(S.E. Est.)**		.307 (.045)	.672 (.031)

*** *p* < .01.**

Regression Coefficients for Final FYGPA Models

Variable	No covariates	With covariates
Item difficulty	--	**.009 (.572)****
Item discrimination	--	**.197 (.393)****
Abstract	**.013 (.222)****	**.007 (.118)***
Non-Routine	**.012 (.135)****	**-.021 (-.244)****
Comprehension	**.011 (.144)****	.004 (.054)
Algebra & Functions	**.011 (.158)****	-.002 (-.035)
Data Analysis	-.008 (-.088)	.001 (.012)
Geometry & Meas.	.006 (.081)	.004 (.051)
Multiple-Choice	**-0.008 (-.128)****	**.007 (.110)****

*** *p* < .05. ** *p* < .01.**

CONCLUSIONS

Abstract vs. Concrete Assessment

- Abstract items were significantly better predictors of FYGPA than concrete items.
- Once item difficulty and discrimination were controlled, the beneficial effect of abstract items on the validity coefficients was reduced but did not disappear.

Cognitive Complexity

- Scores on the items coded at the highest level of cognitive complexity (NONR) were positively related to FYGPA.
- **An unexpected result** – after controlling for difficulty and discrimination, these items actually had significantly lower validity coefficients.
- **Possible explanation** – these items are more difficult, contributing positively to their validity; but their complexity makes them prone to misinterpretation, thus decreasing their correlation with outcomes after difficulty and discrimination are controlled.

Multiple-Choice vs. Constructed Response

- MC items had significantly lower validity coefficients compared to SPR items.
- When item difficulty and discrimination were controlled, MC items had larger correlations with FYGPA.

Content Area

- Provided significant increment to prediction of FYGPA (and math course grade) validity coefficients.
- This effect disappeared in the FYGPA model once item difficulty and discrimination were controlled, but remained in the math course grade model.

Next Steps

- Similar research is underway to examine item characteristics and item-level validity on the SAT critical reading and writing tests.
- Subgroup differences: gender, race/ethnicity, and language.

Thank You!

- College Board researchers are encouraged to freely express their professional judgment. Therefore, the points of view or opinions stated in this presentation do not necessarily represent official College Board position or policy.
- A version of this paper is in press in *Educational and Psychological Measurement*
- Please forward any questions, comments, and suggestions to Jennifer Kobrin at jkobrin@collegeboard.org.

CollegeBoard
inspiring minds™

DELETED SLIDES (FOR REFERENCE IF NEEDED)

SAT-M Items from 12 Forms (n=644)

	N	Percent
Content Area		
Number & operations (NO)	131	20.3
Algebra & functions (AF)	254	39.4
Geometry & measurement (GM)	183	28.4
Data Analysis/stat/probability (DA)	76	11.8
Format		
Multiple-choice (MC)	526	81.7
Student produced response (SPR)	118	18.3
Cognitive Category		
Routine (RTNE)	78	12.1
Comprehension (COMP)	406	63.0
Non-routine/insightful (NONR)	160	24.8
Abstract/Concrete		
Abstract (ABS)	468	72.7
Concrete (CONC)	176	27.3

Change in R-Square Associated with Each Predictor: Math Course Grade Model

Model	Predictors	No covariates	With covariates
1	Cohort	**.084***	**.084***
2	Add Item difficulty/ discrimination	--	**.375***
3	Add Abstract	**.067***	**.032***
4	Add Cognitive	**.033***	**.021***
5	Add Content	**.025***	**.010***
6	Add Format	**.025***	.001
7	Add 2-way interactions	.012	.019
8	Add 3-way interactions	.012	.012
Total R-Sq. **(S.E. Est.)**		.258 (.044)	.555 (.035)

*** *p* < .01.**

Regression Coefficients for Final Math Course Grade Model

Variable	No covariates	With covariates
Item difficulty	--	**.008 (.488)****
Item discrimination	--	**.171 (.353)****
Abstract	**.014 (.244)****	**.009 (.164)****
Non-Routine	**.008 (.096)****	**-.011 (-.128)****
Comprehension	**.010 (.142)****	**.006 (.090)****
Algebra & Functions	**.008 (.124)****	.003 (.050)
Data Analysis	-.005 (-.056)	< .0001 (-.005)
Geometry & Meas.	**.009 (.119)****	**.005 (.073)***
Multiple-Choice	**-.010 (-.159)****	.003 (.040)

*** *p* < .05. ** *p* < .01.**

CollegeBoard
inspiring minds™

Change in R-Square Associated with Each Predictor: STEM Percentage

Model	Predictors	No covariates	With covariates
1	Cohort	**.350***	**.350***
2	Add Item difficulty/ discrimination	--	**.221***
3	Add Abstract	.005	.001
4	Add Cognitive	**.049***	.001
5	Add Content	.006	.000
6	Add Format	**.022***	.001
7	Add 2-way interactions	.016	.018
8	Add 3-way interactions	.017	.014
Total R-Sq. (S.E. Est.)		.465 (.016)	.607 (.013)

*** *p* < .01.**

Regression Coefficients for Final STEM Percentage Model

Variable	No covariates	With covariates
Item difficulty	--	**.003 (.446)****
Item discrimination	--	**.012 (.057)***
Abstract	**.002 (.066)***	<.0001 (.020)
Non-Routine	**.008 (.236)****	.001 (.021)
Comprehension	<-.0001 (-.003)	- .001 (-.038)
Algebra & Functions	**.002 (.090)***	.001 (.028)
Data Analysis	-.001 (-.037)	- .001 (-.019)
Geometry & Meas.	< .0001 (.012)	< .0001 (.006)
Multiple-Choice	**-.004 (-.147)****	- .001 (-.031)

*** $p < .05$. ** $p < .01$.**

CollegeBoard
inspiring minds™