



# Creating More Secure Exams through Performance Based Testing

Andrew Wiley  
The College Board  
Research and Development

**February 25, 2009**



# Background

- Choosing students: Higher education admissions tools for the 21<sup>st</sup> century (Camara & Kimmel, 2005)
- Purpose:
  - Identify additional predictors of college success
  - Expand the definition of what constitutes successful performance in college beyond freshman GPA
- College Board has initiated several projects to address this research area

# Background

- Most of these projects involves the development of measures that are closer to performance based assessments than are the traditional exams like the SAT.
- The challenge that The College Board must face is whether these new assessments can be delivered in a manner that is secure and not easily coached.



## Research collaboration with Michigan State University

- Identify a broader domain of college student performance:
  - Review university mission statements and department objectives
  - Interview with university staff responsible for student life at Michigan State University
  - Review of the education literature on student outcomes
- Our systematic search resulted in 12 dimensions of student performance...

# 12 Dimensions of Student Performance

Broadening the Performance Domain in the Prediction of Academic Success (Schmitt, Oswald, & Gillespie, 2004)

1. Knowledge, learning, mastery of general principles
2. Continuous learning, intellectual interest and curiosity
3. Artistic and cultural appreciation
4. Multicultural appreciation
5. Leadership
6. Interpersonal skills
7. Social responsibility, citizenship and involvement
8. Physical and psychological health
9. Career orientation
10. Adaptability and life skills
11. Perseverance
12. Ethics and integrity

## Two “Noncognitive” Measures

- **Situational judgment inventory**
  - A situation is presented along with several alternative courses of action.
  - The respondent is asked to indicate what she/he would be most likely and least likely to do.
- **Biodata**
  - Short, multiple choice reports of past experience/background and interests/preferences.

# Study 1: Psychometric adequacy & scale refinement

- 644 MSU freshmen completed one of the two parallel forms of the biodata and SJI instruments at the beginning of the academic year.
- Identical empirical-keying procedures were conducted on both instruments at the item level (double-cross validated using randomly split samples).
- Results indicated significant incremental validity for some of the scales above and beyond the validity of SAT/ACT scores and existing measures of personality in predicting college GPA.
- The biodata and SJI demonstrated the greatest incremental validity when absenteeism, students' self ratings, and peer-ratings of performance were examined ( .19, .22, and .14, respectively).

## Study 1: Standardized Differences Compared with White group...

Non-cognitive Dimension	Black	Hispanic	Asian
Knowledge	-0.08	-0.20	-0.25
Learning	0.01	0 .63*	-0.19
Artistic	-0.19	0 .73*	0.15
Multicultural	-0.11	0 .63*	0.02
Leadership	-0.18	0.08	-0.30
Interpersonal	-0.18	0.33	-0.38*
SJI composite	-0.05	-0.14	-0.21
Citizenship	0.05	0.23	-0.14
Health	-0.31*	0.06	-0.67*
Career	0 .34*	0 .56*	0.14
Adaptability	0.03	0.09	-0.41*
Perseverance	0.13	0 .55*	-0.18
Ethics	0.17	-0.06	-0.13

- Positive values indicate that minorities perform **better** than White students.
- The *d* values for biodata and SJI measures across ethnic and gender subgroups were consistently smaller than those found on cognitive predictors.
- \*  $p < .05$



# Study 2: Predicting FYGPA: Total Sample across 10 Institutions ( $N = 2443$ )

Variable	Mean	SD	Validity	Regression Wt.
HS-GPA	3.51	.42	.61	.70*
SAT/ACT	.61	.91	.64	.38*
<b><i>R (Adjusted R)</i></b>				<b>.70 (.70)</b>
<u>Biodata</u>				
Knowledge	3.15	.48	.29	.08*
Learning	3.08	.61	.13	-.00
Artistic	2.91	.83	.22	.02
Diversity	2.98	.66	.13	.03
Leadership	3.07	.81	.14	-.02
Responsibility	3.35	.76	.19	.02
Health	3.26	.51	.16	.10*
Citizenship	3.31	.65	-.17	-.15*
Adaptability	3.38	.44	.11	-.02
Perseverance	3.73	.49	.09	.08*
Ethics	3.86	.55	.18	-.01
<u>SJI Composite</u>	.67	.33	.27	.23
Constant				-.28*
<b><i>R (Adjusted R)</i></b>				<b>.72 (.72)</b>
<b>Change in <i>R</i></b>				<b>.02*</b>



## Predicting Self-Rated Performance: Total Sample across 10 Institutions ( $N = 900$ )

Variable	Mean	SD	Validity	Regression Wt.
HS-GPA	3.64	.35	.09*	.03
SAT/ACT	.98	.81	.03	-.02
<b><i>R (Adjusted R)</i></b>				<b>.09 (.08)</b>
<u>Biodata</u>				
Knowledge	3.23	.46	.25*	.01
Learning	3.16	.61	.26*	.02
Artistic	3.08	.85	.25*	.05*
Diversity	3.05	.69	.30*	.07*
Leadership	3.11	.80	.33*	.07*
Responsibility	3.42	.74	.31*	.04
Health	3.28	.48	.21*	.10*
Citizenship	3.22	.69	.23*	.07*
Adaptability	3.42	.44	.26*	.12*
Perseverance	3.74	.46	.32*	.04
Ethics	3.92	.51	.27*	.10*
<u>SJI Composite</u>	.73	.30	.28*	.16
Constant				1.12*
<b><i>R (Adjusted R)</i></b>				<b>.50* (.49)</b>
<b>Change in <i>R</i></b>				<b>.41*</b>



# Predicting Class Absenteeism: Total Sample across 10 Institutions ( $N = 899$ )

Variable	Mean	SD	Validity	Regression Wt.
HS-GPA	3.64	.35	-.04	.34*
SAT/ACT	.98	.81	.17*	.36*
<b><i>R</i> (Adjusted <i>R</i>)</b>				<b>.22* (.22)</b>
<u>Biodata</u>				
Knowledge	3.23	.46	-.15*	-.14
Learning	3.16	.61	-.06*	-.04
Artistic	3.09	.85	-.03	-.03
Diversity	3.05	.69	-.03	.03
Leadership	3.11	.80	-.05	.05
Responsibility	3.42	.74	-.08*	-.02
Health	3.27	.48	-.18*	-.34*
Citizenship	3.22	.69	-.06*	.11*
Adaptability	3.42	.44	-.09*	.08
Perseverance	3.74	.46	-.17*	-.14
Ethics	3.92	.51	-.19*	-.18*
<u>SJI Composite</u>	.73	.29	-.17*	.41*
Constant				5.17*
<b><i>R</i> (Adjusted <i>R</i>)</b>				<b>.36* (.34)</b>
<b>Change in <i>R</i></b>				<b>.14*</b>



# Representative Subgroup Differences in Standardized Units

*Compared with White group*

	SAT/ACT	HS-GPA	SJI	Persevere	Career	Learn	Responsible
Hispanic	-.83	-.61	-.18	-.01	.08	-.05	.00
Asian	.38	.12	-.03	-.12	-.03	-.18	.07
African-American	-1.15	-.81	-.23	.20	.52	-.11	-.14

*Compared with Male group*

	SAT/ACT	HS-GPA	SJI	Persevere	Career	Learn	Responsible
Females	-.42	.02	.36	.22	.17	-.20	.26

< .20 = small effect  
 .20-.50 = moderate effect  
 > .50 = large effect



## Percent of Students Selected: Two Composites and Three Selection Strategies

Group	Top 85%		Top 50%		Top 15%	
	AB	AB+	AB	AB+	AB	AB+
Hispanic	4.4 →	4.6 (+.2)	4.1 →	4.9 (+.8)	3.9 →	5.5 (+1.6)
Asian	7.6 →	7.7 (+.1)	9.9 →	9.5 (-.4)	17.5 →	12.9 (-4.6)
African-American	17.9 →	19.8 (+1.9)	9.6 →	13.6 (+4.0)	1.3 →	7.2 (+5.9)
White	70.2 →	67.9 (-2.3)	76.4 →	71.9 (-4.5)	77.2 →	74.4 (-2.8)

**AB** = equally weighted composite of HSGPA and SAT/ACT.  
**AB+** = equally weighted composite of HSGPA, SAT/ACT, Biodata, and SJI.

# Limitations & Future Research

- Public relations and acceptance of these measures by consumers (i.e., admissions officers, parents, students). Need to collect reactions to new admissions measures along a variety of dimensions (e.g., fairness, face validity).
- Fakability in high-stakes situation especially relevant for biodata, less so for SJI. However, note that essays can be coached and edited, and self-reported activities can also be inflated.
- More research and evaluation efforts need to be conducted when these measures are used operationally in college settings.

## Study 3: Purpose & Research Questions

- **Purpose:** evaluating the utility of the biodata and situational judgment measures in as close to a real admissions situation as is possible
  - Administer new measures to college applicants rather than college freshmen.
  - On an annual basis, collect class absenteeism, self rated performance of the noncognitive dimensions, and commitment to the university from enrolled students; institutions will provide course grades and retention information.
- **Research Questions:**
  - The incremental validity of the biodata and the situational judgment measures will be assessed after controlling for high school GPA and SAT/ACT scores.
  - Differential prediction will also be assessed to see whether each measure-outcome relationship differs across various subgroups (e.g., gender and race).
  - The relationship between scores on these noncognitive measures and holistic file review will be examined to test whether these measures could be substituted for the more subjective file review.



## Preliminary Validity Results...

- A year prior to Study 3 data collection, a similar pilot study was conducted with only Michigan State University applicants.
- Comparisons between this sample and our past studies should reveal the degree to which the application process itself affects mean scores, variability, reliability, and validity of these scales.



# MSU Pilot: Demographic Statistics

Variable	Predictor		Outcome	
	N	%	N	%
<b>Ethnic Status</b>				
Hispanic	25	4.5	5	4.0
Asian	25	4.5	3	2.4
African American	19	3.4	0	0.0
Caucasian		463	83.1	107
84.9				
Other	25	4.5	11	8.8
<b>Gender</b>				
Male	215	37.6	41	32.5
Female	357	62.4	83	65.9

Note. For Ethnic Status, the Hispanic group includes respondents of Mexican, Puerto Rican, and Hispanic origin. Total sample size varies across the demographic categories due to missing data. Response categories for major varied across the two data collections.

# MSU Pilot: Results – Mean Differences

<b>Dimensions</b>	<b>Average score at MSU 2006-2007</b>	<b>Average score all 10 universities 2004</b>	<b>d-value</b>
Knowledge	3.41 (.46)	3.15 (.47)	<b>.54</b>
Continuous Learning	3.40 (.62)	3.09 (.61)	<b>.50</b>
Artistic Appreciation	3.15 (.78)	2.91 (.82)	<b>.29</b>
Multicultural Appreciation	3.25 (.66)	2.98 (.66)	<b>.41</b>
Leadership	3.35 (.77)	3.07 (.81)	<b>.35</b>
Social Responsibility	3.67 (.70)	3.32 (.76)	<b>.46</b>
Health	3.40 (.51)	3.25 (.51)	<b>.30</b>
Career Orientation	3.45 (.61)	3.32 (.65)	<b>.20</b>
Adaptability	3.49 (.46)	3.38 (.45)	<b>.24</b>
Perseverance	3.88 (.47)	3.73 (.49)	<b>.31</b>
Ethics	4.13 (.46)	3.86 (.54)	<b>.52</b>
Jobs Scale	2.51 (.86)	2.80 (.58)	<b>-.26</b>
Awards Scale	2.24 (.69)	2.42 (.70)	<b>-.29</b>
<i>SJI</i>	.42 (.14)	.33 (.17)	<b>.56</b>

Note. Standard deviations are in parentheses next to the means. Positive d values indicate that the 2007 applicant sample had scores higher than the 2004 student sample.



# Incremental Validity of Biodata Measures

Outcomes	N	$R^2$ (HSGPA, SAT)	Overall $R^2$	$\Delta R^2$
BARS	57	0.023	0.443*	0.420*
OCB	57	0.017	0.392	0.374*
Deviance	57	0.025	0.373	0.348
Turnover Intent	58	0.077	0.248	0.172
Academic Satisfaction	58	0.008	0.353	0.345
Social Satisfaction	58	0.077	0.294	0.218
FYGPA	84	0.201*	0.335*	0.134
Absenteeism	58	0.061	0.234	0.173

- To preserve N in these regressions, the SJI was not included because of a relatively low response rate to this measure.
- It is worth noting that small sample sizes, such as those observed in these analyses, can seriously limit the ability to detect significant relationships due to decreased statistical power.



# Thank You

Thanks to ATP

and

Thanks to you



# 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:  
Andrew Wiley at: [Awiley@collegeboard.org](mailto:Awiley@collegeboard.org)