The International Cognitive Ability Resource: Development and initial validation of a public-domain measure

David M. Condon
William Revelle

Thirteenth Annual ISIR Conference
San Antonio, Texas
December 15, 2012

International Cognitive Ability Resource ("ICAR")

International Cognitive Ability Resource ("ICAR")

- A *new* measure of cognitive ability?

International Cognitive Ability Resource ("ICAR")

- A *new* measure of cognitive ability?
- Why bother?

International Cognitive Ability Resource ("ICAR")

- A *new* measure of cognitive ability?
- Why bother?
- None of the extant measures serve our research needs

  ...and, as we've discovered, other research groups share this problem.

ICAR: How does it differ from existing measures?

1. A **public-domain** measure
   - More convenient, affordable, and flexible administration —> maximally *reproducible* data

2. Not confined to controlled environments (i.e., can be administered over the internet)
   - unproctored
   - power items
   - "Google"-resistant content
   - draw heavily on automatic item generation techniques

   ...though all nature of item types can be included, regardless of development/administration methods.

3. (quasi) Open-source development and distribution
   - collaboratively developed & maintained by the researchers who use it

ICAR: Two common critiques

1. "You're giving up the keys to the kingdom"
   - Designed for use in research contexts – not a substitute for proprietary measures used in clinical/diagnostic settings
   - Not recommended for use in selection or high-stakes assessment (though this might someday be possible)

2. "Copyrights are necessary to maintain sufficient validity"
   - The pace of scientific research may be diminished by reliance on proprietary measures (Goldberg, 1999)
   - Copyrights address the prospect of item disclosure by:
     – reducing transparency about item content
     – making testing more difficult
   - An alternative is to *decrease the harm* caused by item disclosure at the stage of item development using automatic item generation.

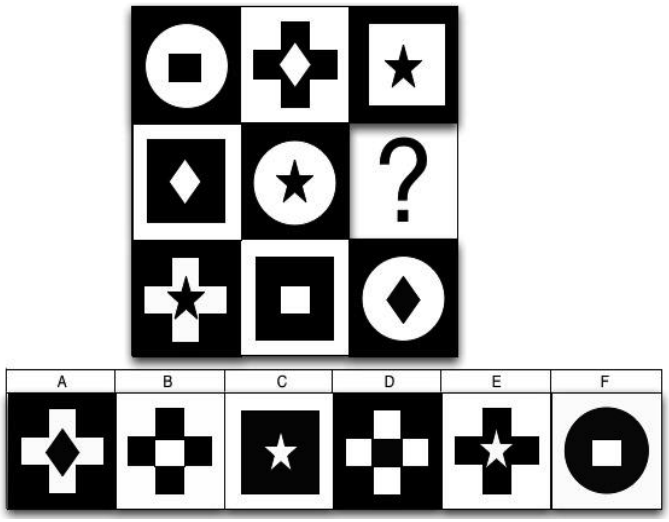
Empirical evaluation of existing ICAR measures

- This is not merely theory.
- We have administered public domain items to 200k+ participants
- Full measure reported on here includes 60 items administered in quasi-random subsets to 35k/yr
- Four existing item types include:
  - Matrix Reasoning items (11 items)
  - Verbal Reasoning items (16 items)
  - Letter and Number Series items (9 items)
  - Three-Dimensional Rotation items (24 items)

– Summarize findings with regards to reliability, structure and validity

Sample ICAR items

Matrix Reasoning



Verbal Reasoning

What number is one fifth of one fourth of one ninth of 900?

(1) 2 (2) 3 (3) 4 (4) 5 (5) 6 (6) 7

If the day after tomorrow is two days before Thursday, then what day is it today?

(1) Friday (2) Monday (3) Wednesday

(4) Saturday (5) Tuesday (6) Sunday

Letter and Number Series

In the following alphanumeric series, what letter comes next?

I J L O S

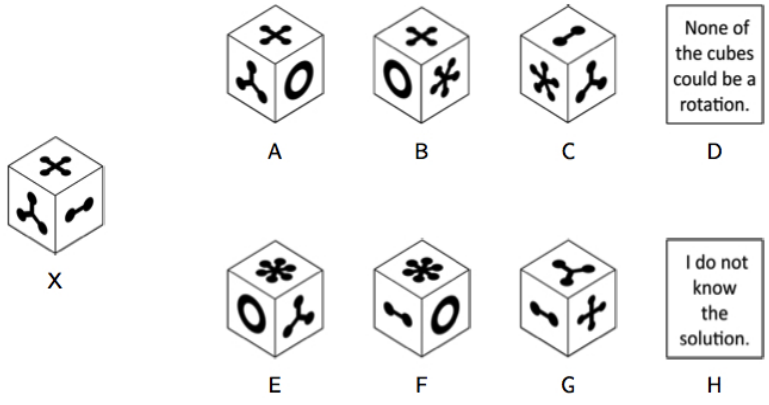
(1) T (2) U (3) V (4) X (5) Y (6) Z

In the following alphanumeric series, what letter comes next?

Q S N P L

(1) J (2) H (3) I (4) N (5) M (6) L

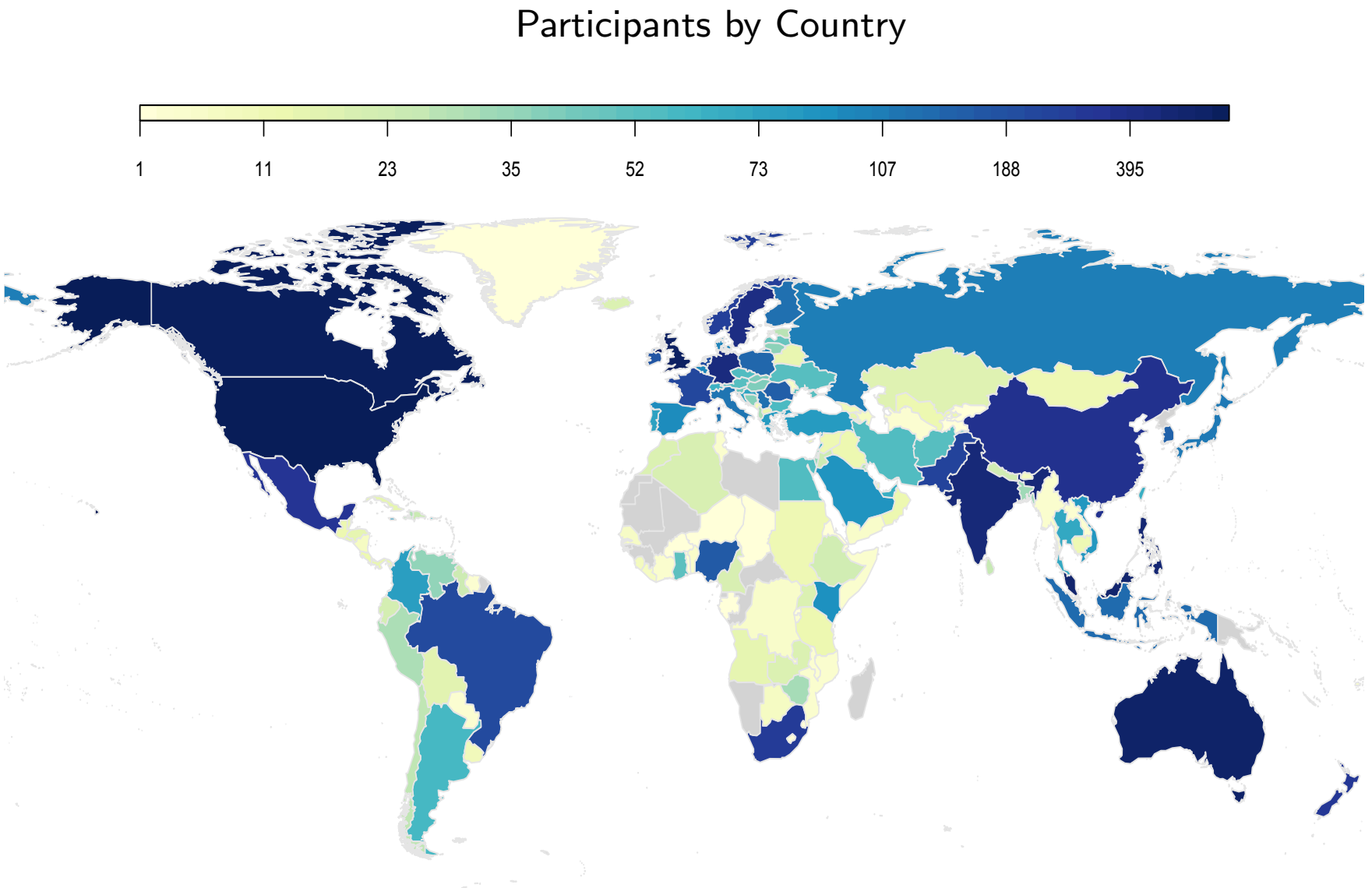
Three-Dimensional Rotation



Three studies summarized in the results

- **Study 1:** Random subsets of 14 to 16 items from the full measure (ICAR60) administered to a large online sample
- Study 2: The 16 item ICAR Sample Test (ICAR16) administered to a subset of the online sample
- Study 3: The 16 item ICAR Sample Test (ICAR16) administered to an offline university sample

Study 1: 80,000 Participants (8/10 - 12/12)

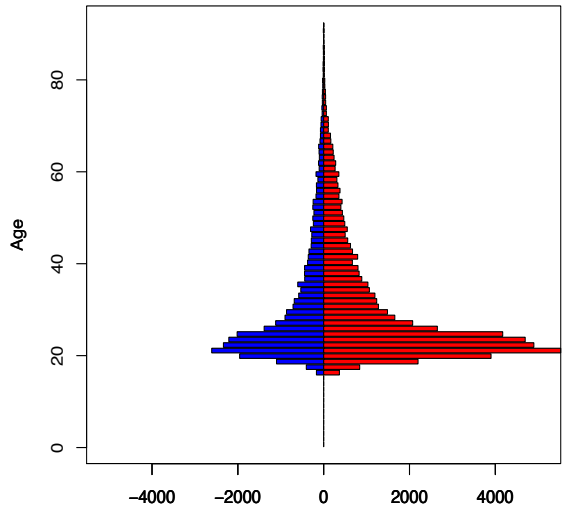


Study 1: 80,000 Participants (8/10 - 12/12)

Participants by Country

Country	Participants
USA	61857
Canada	3691
United Kingdom	1861
Australia	1465
Malaysia	1421
Philippines	816
India	807
Germany	525
Sweden	395
Singapore	338

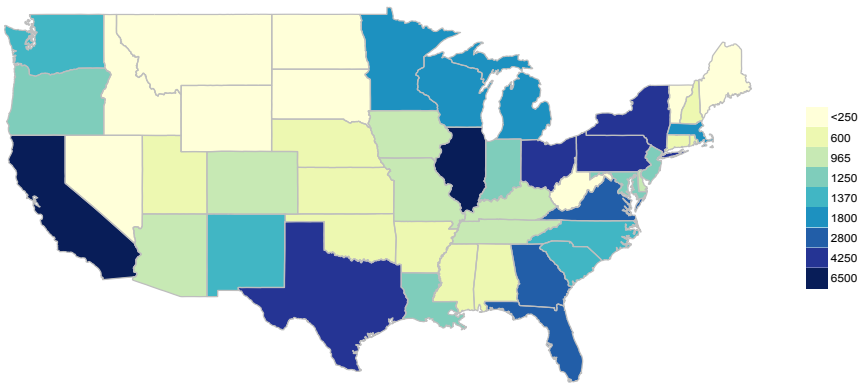
Age by Males and Females



Age: m=26.1, sd=10.7, med=22

Gender: 66.4% female

76.9% of total from U.S.



Ethnicity	% of U.S.
White	67.8
African-American	10.4
Hispanic-American	7.8
Two or more	6.2
Asian-American	4.7
Native American	1.2
Other	1.8

Three studies summarized in the results

- Study 1: Random subsets of 14 to 16 items from the full measure (ICAR60) administered to a large online sample (80k)
- **Study 2:** The 16 item ICAR Sample Test (ICAR16) administered to a subset of the online sample
  - 4 items of each type
  - 1,909 university-age participants (age: $m = 19.7$ yrs, $sd = 1.4$; 72% female)
- Study 3: The 16 item ICAR Sample Test (ICAR16) administered to an offline university sample

Three studies summarized in the results

- Study 1: Random subsets of 14 to 16 items from the full measure (ICAR60) administered to a large online sample (80k)
- Study 2: The 16 item ICAR Sample Test (ICAR16) administered to a subset of the online sample
  - 4 items of each type
  - 1,909 university-age participants (age: $m = 19.7$ yrs, $sd = 1.4$; 72% female)
- **Study 3:** The 16 item ICAR Sample Test (ICAR16) administered to an offline university sample
  - 16 item ICAR Sample Test and *Shipley-2* Composites A and B
  - 137 student participants (age: $m = 19.7$ yrs, $sd = 1.2$; 55% female)

Results: Reliability of the ICAR measures

The full 60 item measure based on composite correlations (n = 80k):

	Items	α	ω_h	ω_t
ICAR60	60	0.93	0.60	0.94
Letter Number Series	9	0.77	0.66	0.80
Matrix Reasoning	11	0.67	0.56	0.70
3D Rotation	24	0.93	0.70	0.95
Verbal Reasoning	16	0.76	0.63	0.77

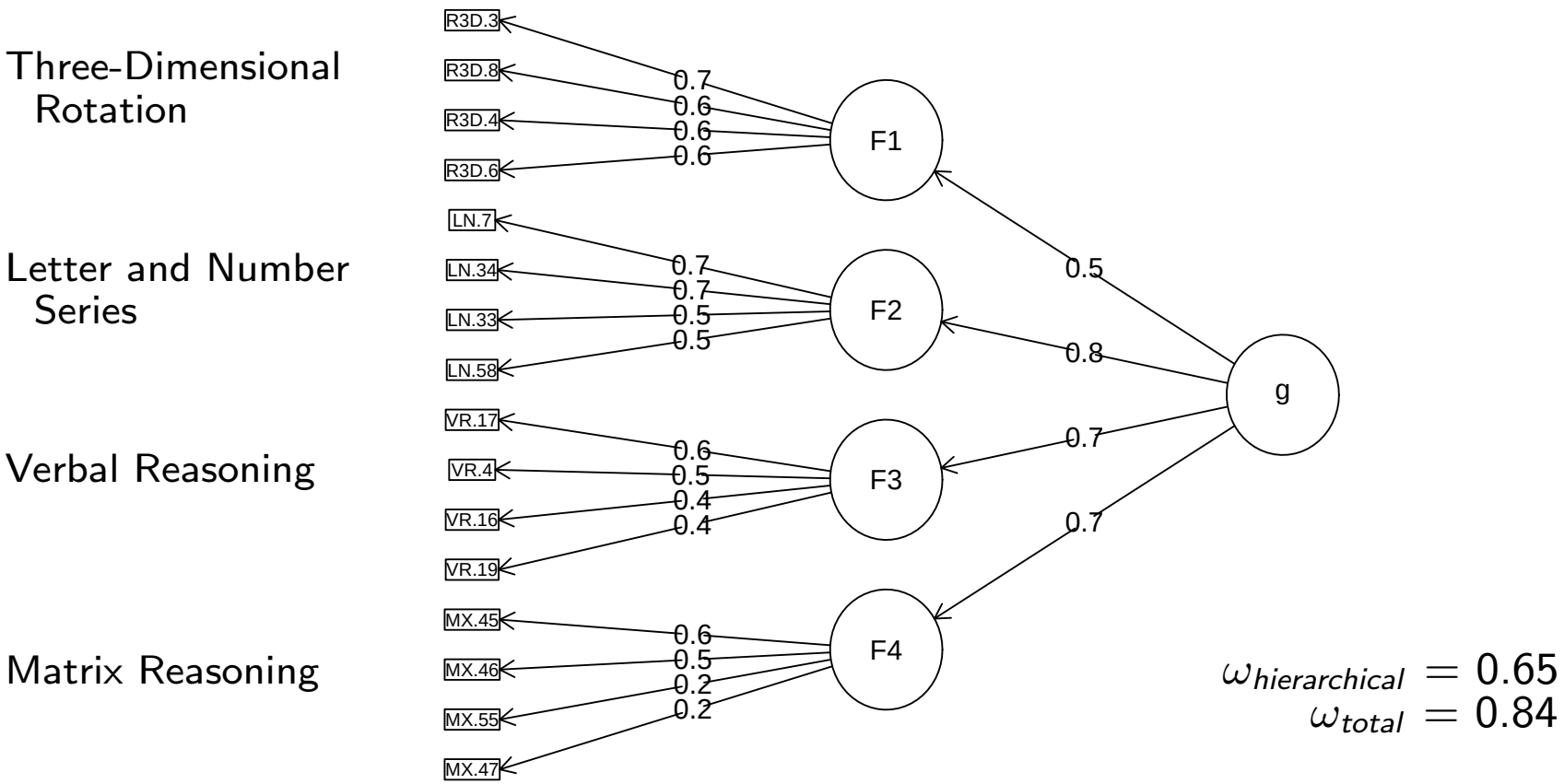
The 16 item ICAR Sample Test:

		University Sample (n = 137)			Online Sample (n = 1909)		
	Items	α	ω_h	ω_t	α	ω_h	ω_t
ICAR16	16	0.76	0.50	0.80	0.81	0.63	0.83
Letter Number Series	4	0.68	0.62	0.82	0.68	0.67	0.71
Matrix Reasoning	4	0.54	0.50	0.60	0.52	0.49	0.56
3D Rotation	4	0.77	0.71	0.80	0.74	0.72	0.76
Verbal Reasoning	4	0.36	0.44	0.48	0.59	0.57	0.62
Shipley - Vocabulary	33	0.61	0.24	0.66			
Shipley - Block Patterns	23	0.83	0.50	0.88			
Shipley - Abstraction	15	0.37	0.45	0.51			

Notes: α = Cronbach's alpha, ω_h = omega hierarchical, ω_t = omega total. Reliabilities calculated on Pearson correlations.

Results: Structural Characteristics of the ICAR Sample Test

Hierarchical factor analysis



Notes: $\omega_{hierarchical}$ = general factor saturation of the model; ω_{total} = total reliable variance

Results: Participant-level correlations with achievement tests

The full 60 item measure (composite):

	Uncorrected				
	SATV	SATQ	SATW	SATVQ	ACT
ICAR60	0.42	0.50	0.37	0.50	0.42
Letter Number	0.31	0.38	0.27	0.37	0.29
Matrix Reasoning	0.27	0.34	0.23	0.33	0.26
3D Rotation	0.31	0.41	0.27	0.39	0.35
Verbal Reasoning	0.52	0.51	0.47	0.55	0.45

	Corrected for reliability				
	SATV	SATQ	SATW	SATVQ	ACT
ICAR60	0.47	0.55	0.41	0.57	0.45
Letter Number	0.38	0.46	0.33	0.47	0.34
Matrix Reasoning	0.36	0.44	0.30	0.45	0.33
3D Rotation	0.35	0.45	0.30	0.45	0.37
Verbal Reasoning	0.64	0.62	0.57	0.70	0.53

Participant level, IRT-based scores:

	Corrected for incidental selection & reliability				
	SATV	SATQ	SATW	SATVQ	ACT
ICAR60	0.47	0.45	0.41	0.50	0.42

Note: All values significant at $p < .001$

Results: Group-level correlations between GRE and ICAR

	ICAR60	Letter Number Series	Matrix Reasoning	3D Rotation	Verbal Reasoning
GREV	0.54	0.45	0.43	0.48	0.65
GREQ	0.77	0.77	0.77	0.82	0.66
GREVQ	0.86	0.82	0.81	0.87	0.82

Notes: All values significant at $p < .001$

GRE scores are group means:

- $N = 569{,}000$ "senior and non-enrolled college graduates" (Educational Testing Service, 2010)
- Took test between July 1, 2005 to June 30, 2008
- 287 "intended graduate major" choices offered with GRE
- Consolidated w/ sample size weighting to 147 university major choices in ICAR
- Correlations based on **88 ICAR majors with more than 50 participants**

Results: Group-level correlations between SAT and ICAR

	ICAR60	Letter Number Series	Matrix Reasoning	3D Rotation	Verbal Reasoning
SATV	0.58	0.51	0.30*	0.50	0.70
SATQ	0.82	0.82	0.73	0.80	0.78
SATVQ	0.77	0.73	0.56	0.71	0.81

Notes: * not significant. All other values significant at $p < .01$

SAT scores are group means:

- $N = 1{,}411{,}595$ "college-bound seniors in class of 2012" (College Board, 2012)
- 38 intended college major choices
- consolidated our 147 university major choices into 31 choices offered w/the SAT
- 7 incompatible majors representing 1.3% of SAT test-takers (exclusive of 9.0% undecided and other)

Results: Correlations with the Shipley-2 (0.87 - 0.99)

Based on participant-level scores in the university sample:

	SATV	SATQ	SATW	ACT	ShipCompA	ShipCompB
Shipley-2 Composite A	0.99	0.95	0.99	0.93	NA	NA
Shipley-2 Composite B	0.90	0.90	0.84	0.89	NA	NA
ICAR16	0.84	0.79	0.76	0.75	0.99	0.87

Correlations corrected for restriction of range and reliability.

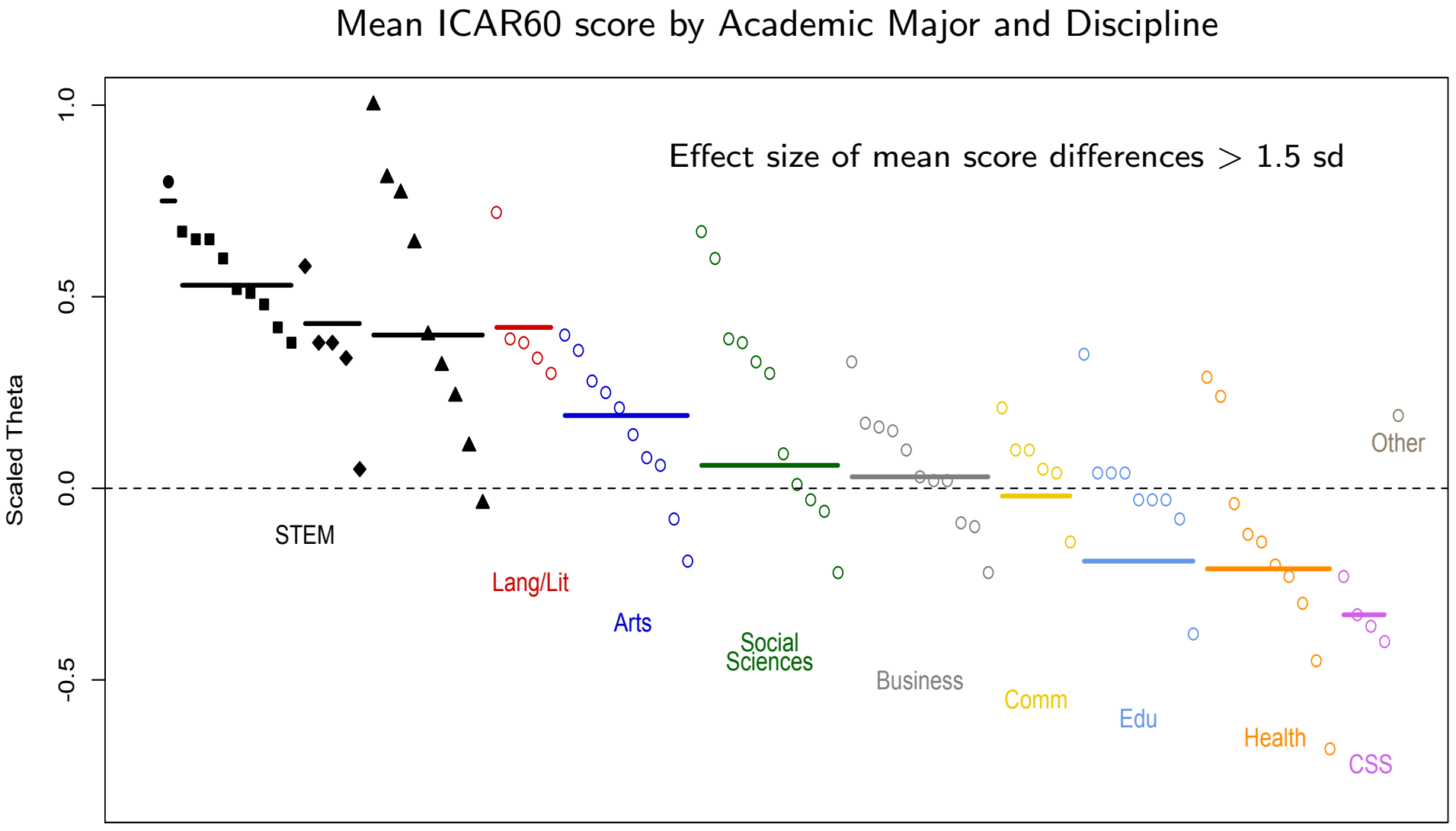
Summary of Results

- Reliability for the full measure based on composite scores is high ($\alpha > 0.9$)
- Factor structure suggests four distinct but correlated factors.
- Corrected correlations of full measure with self-reported achievement test scores range from 0.50 - 0.57 in the online sample to 0.75 - 0.84 in the university sample.
- Group level correlations based on major were 0.77 - 0.86 between the full measure and achievement test scores.
- Corrected correlations for the 16 item ICAR Sample Test with the *Shipley-2* 0.87 - 0.99.

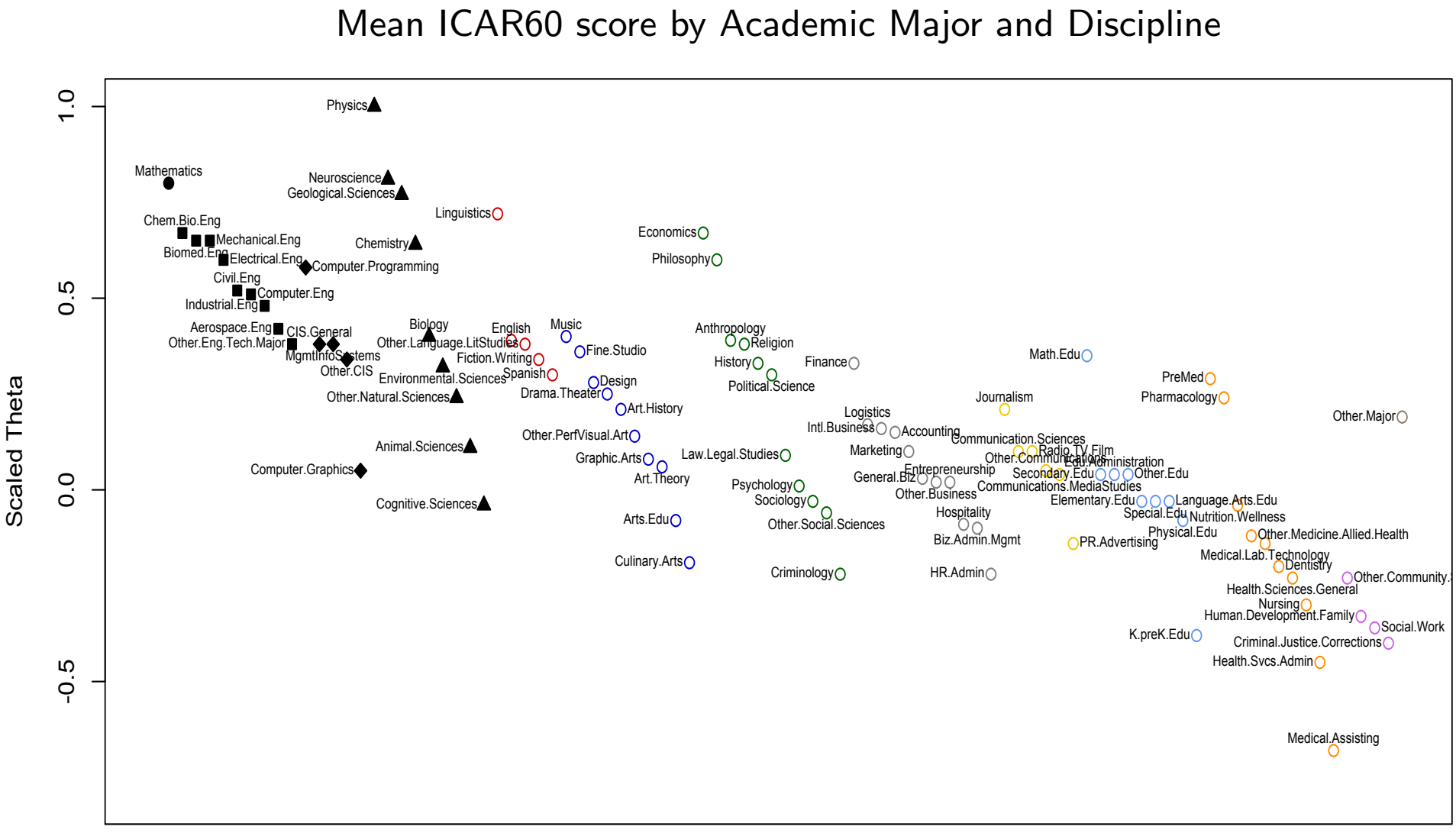
Future Directions: Further development underway

- Focus on automatic item generation techniques
- Broaden scope of item types:
  - Exploring use of "cloze"-type reading comprehension items to assess verbal ability
  - Several spatial ability item types under consideration or development.
    - 2D rotation tasks
    - map-reading/navigation
    - paper folding
    - cross-section of 3D objects (Cohen & Hegarty, 2007)

Future Directions: Applications in individual differences research



Future Directions: Applications in individual differences research



Summary

- Validation work is never done but we are encouraged by the findings thus far.
- Collaboration will be the key to rapid development and adoption of the measure among the intelligence research community.

Supplementary Materials

TAI Model of Individual Differences

Synthetic Aperture Personality Assessment ("SAPA")
- cross-sectional, large-scale assessment over the Internet

- **Temperament**
  $>$ 2,400 public-domain IPIP items
  $>$ 1,350 non-proprietary non-IPIP items
  50+ trait constructs evaluated since 2008
- **Abilities**

- **Interests**
  8 public-domain vocational scales based on Holland's RAISEC
  33 public-domain avocational scales

SAPA Administration to 80,000 Participants (8/10 - 12/12)

- SAPA methodology:
  - Administer subset of items to each participant and create synthetic correlation matrices.
  - 125 unique participants each day
  - Participation driven by response-based feedback on temperament
    - 20 demographic variables
    - 60 temperament and interest items
    - 16 ICAR items
  - Across participants, administering 200-600 items at a time.
  - Efficient exploration of item-level correlations within and between scales.

Results: Validity

The full 60 item measure (composite):

	SATV	SATQ	SATW	ACT	ICAR60	ICAR-LN	ICAR-MR	ICAR-R3D	ICAR-VR
SATV	0.86	0.84	0.97	0.65	0.47	0.38	0.36	0.35	0.64
SATQ	0.73	0.88	0.82	0.63	0.55	0.46	0.44	0.45	0.62
SATW	0.84	0.72	0.88	0.63	0.41	0.33	0.30	0.30	0.57
ACT	0.59	0.58	0.58	0.95	0.45	0.34	0.33	0.37	0.53
ICAR60	0.42	0.50	0.37	0.42	0.93	0.92	0.98	0.96	0.94
ICAR-LN	0.31	0.38	0.27	0.29	0.78	0.77	0.88	0.60	0.92
ICAR-MR	0.27	0.34	0.23	0.26	0.77	0.63	0.67	0.68	0.84
ICAR-R3D	0.31	0.41	0.27	0.35	0.89	0.51	0.54	0.93	0.59
ICAR-VR	0.52	0.51	0.47	0.45	0.79	0.70	0.60	0.50	0.76

Uncorrected correlations below the diagonal, correlations above the diagonal corrected for reliability.

Participant level, IRT-based scores:

	SATV	SATQ	SATW	ACT	ICAR60
SATV	0.86	0.87	0.97	0.69	0.47
SATQ	0.73	0.88	0.82	0.64	0.45
SATW	0.84	0.72	0.88	0.66	0.41
ACT	0.59	0.58	0.58	0.95	0.42
ICAR60	0.33	0.39	0.3	0.33	0.93

Uncorrected correlations below the diagonal, correlations above the diagonal corrected for incidental selection effects and reliability.

Validity: Uncorrected correlations

Comparison of correlations using different scoring methods

Using composite scale scores

	SATV	SATQ	SATW	SATVQ	SATVQW	ACT	ICAR60	LN	MX	R3D
ICAR60	0.42	0.50	0.37	0.50	0.46	0.42				
ICAR-LN	0.31	0.38	0.27	0.37	0.33	0.29	0.78			
ICAR-MX	0.27	0.34	0.23	0.33	0.30	0.26	0.77	0.63		
ICAR-R3D	0.31	0.41	0.27	0.39	0.35	0.35	0.89	0.51	0.54	
ICAR-VR	0.52	0.51	0.47	0.55	0.53	0.45	0.79	0.70	0.60	0.50

Notes: All values are statistically significant at $p < .001$

Using IRT-based scoring

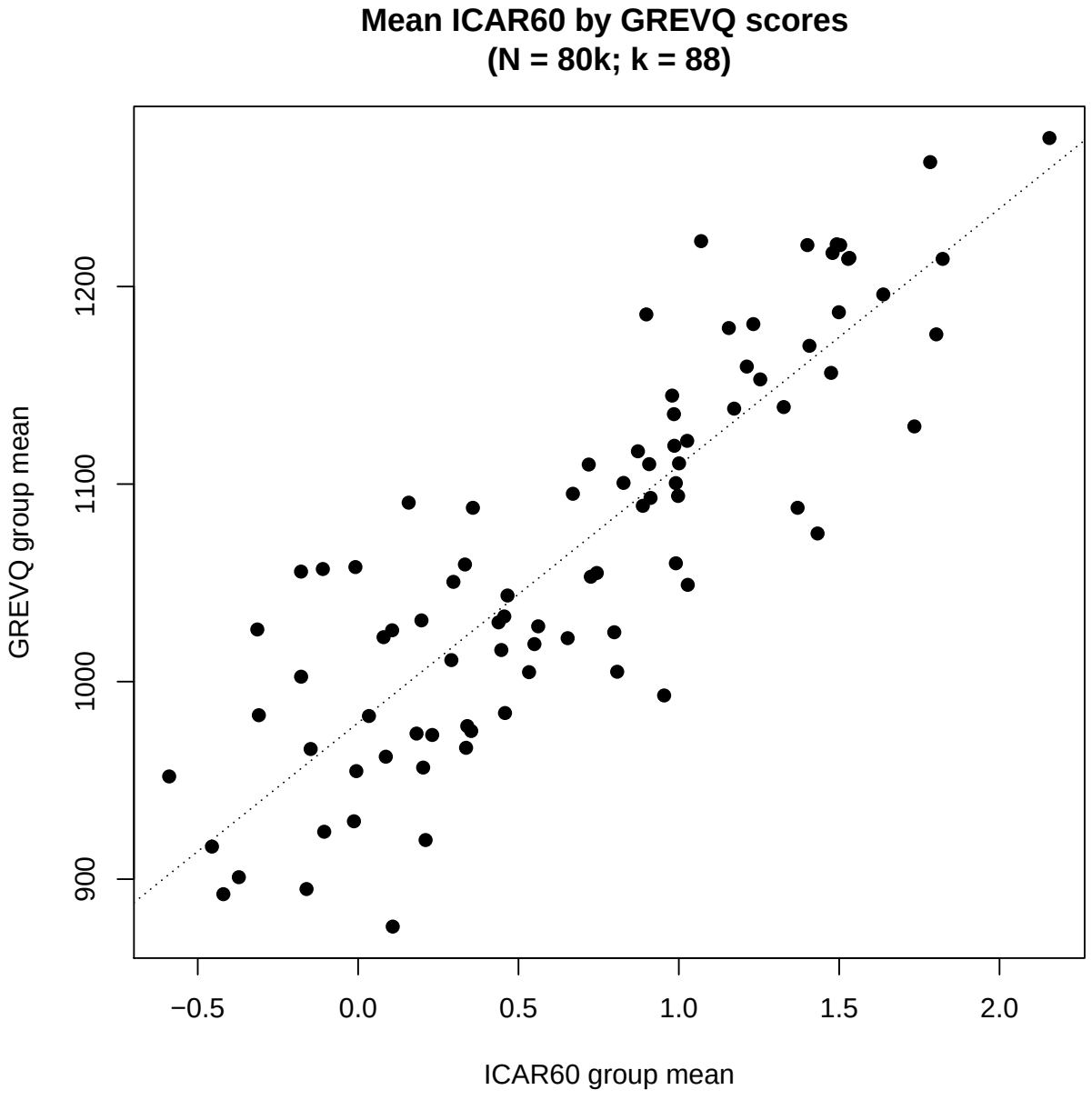
	SATV	SATQ	SATW	SATVQ	SATVQW	ACT	ICAR60	LN	MX	R3D
ICAR60	0.33	0.39	0.30	0.38	0.36	0.33				
ICAR-LN	0.22	0.27	0.20	0.26	0.24	0.21	0.70			
ICAR-MX	0.16	0.21	0.15	0.20	0.18	0.16	0.56	0.29		
ICAR-R3D	0.22	0.29	0.19	0.27	0.25	0.24	0.65	0.26	0.23	
ICAR-VR	0.33	0.33	0.30	0.35	0.34	0.32	0.69	0.38	0.26	0.23

Notes: All values are statistically significant at $p < .001$

Validity: About corrections

- Two corrections are warranted for the correlations with achievement test scores:
  - Correction for (imperfect) reliability
    - For achievement test scores, using meta-analytic findings of actual-to-self-report correlations (Kuncel, Crede & Thomas, 2005; Mayer, Stull, Campbell, Almeroth, Bimber, Chun & Knight, 2006; Cole & Gonyea, 2009)
  - Correction for incidental selection effect caused by optional self-reporting of achievement test scores
    - Need to correct for an unidentified and unmeasured variable(s) influencing score-reporting
    - Using the two-step "Heckman" correction method (Heckman, 1976, 1979; Greene, 2008; Toomet & Henningsen, 2008)
- Note that correction for range restriction is not warranted in the online sample.

Validity – Group-level correlations between GRE and ICAR



Cohen, C. A. & Hegarty, M. (2007). Sources of difficulty in imagining cross section of 3d objects. *conference paper*, 1–6.

Cole, J. S. & Gonyea, R. M. (2009). Accuracy of self-reported SAT and ACT test scores: Implications for research. *Research in Higher Education*, *51*(4), 305–319.

College Board (2012). *2012 college-bound seniors total group profile report*. New York.

Educational Testing Service (2010). Table of GRE scores by intended graduate major field.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. D. Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe*. Tilburg, The Netherlands: Tilburg University Press.

Greene, W. H. (2008). *Econometric Analysis* (6th Edition ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In S. V. Berg (Ed.), *Annals of Economic and Social Measurement, Volume 5, number 4* (pp. 475–492). Cambridge, MA: NBER.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*(1), 153–161.

Kuncel, N. R., Crede, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, *75*(1), 63–82.

Mayer, R. E., Stull, A. T., Campbell, J., Almeroth, K., Bimber, B., Chun, D., & Knight, A. (2006). Overestimation bias in self-reported SAT scores. *Educational Psychology Review*, *19*(4), 443–454.

Toomet, O. & Henningsen, A. (2008). Sample selection models in R: Package sampleSelection. *Journal of statistical software*, *27*(7), 1–23.