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

David M. Condon*, William Revelle

Northwestern University, Evanston, IL

Abstract

For all of its versatility and sophistication, the extant toolkit of cognitive ability measures lacks a public-domain method for large-scale, remote data collection. While the lack of copyright protection for such a measure poses a theoretical threat to test validity, the effective magnitude of this threat is unknown and can be offset by the use of modern test-development techniques. To the extent that validity can be maintained, the benefits of a public-domain resource are considerable for researchers, including: cost savings; greater control over test content; and the potential for more nuanced understanding of the correlational structure between constructs. The International Cognitive Ability Resource was developed to evaluate the prospects for such a public-domain measure and the psychometric properties of the first four item types were evaluated based on administrations to both an offline

*Correspondence concerning this article should be addressed to David M. Condon, Department of Psychology, Northwestern University, Evanston IL, 60208. Telephone number: 847-491-4515. Email: davidcondon2009@u.northwestern.edu. With thanks to Melissa Mitchell.
Please do not cite without permission.

university sample and a large online sample. Concurrent and discriminative validity analyses suggest that the public-domain status of these item types did not compromise their validity despite administration to 97,000 participants. Further development and validation of extant and additional item types is recommended.

Keywords: cognitive ability, intelligence, online assessment, psychometric validation, public-domain measures, spatial reasoning, matrix reasoning

1. Introduction

The domain of cognitive ability assessment is now populated with dozens, possibly hundreds, of proprietary measures (Camara et al., 2000; Carroll, 1993; Cattell, 1943; Eliot and Smith, 1983; Goldstein and Beers, 2004; Murphy et al., 2011). While many of these are no longer maintained or administered, the variety of tests in active use remains quite broad, providing those who want to assess cognitive abilities with a large menu of options. In spite of this diversity, however, assessment challenges persist for researchers attempting to evaluate the structure and correlates of cognitive ability. We argue that it is possible to address these challenges through the use of well-established test development techniques and report on the development and validation of an item pool which demonstrates the utility of a public-domain measure of cognitive ability for basic intelligence research. We conclude by imploring other researchers to contribute to the on-going development, aggregation and maintenance of many more item types as part of a broader, public-domain tool – the International Cognitive Ability Resource ("ICAR").

2. The Case For A Public Domain Measure

To be clear, the science of intelligence has historically been well-served by commercial measures. Royalty income streams (or their prospect) have encouraged the development of testing "products" and have funded their ongoing production, distribution and maintenance for decades. These assessments are broadly marketed for use in educational, counseling and industrial contexts and their administration and interpretation is a core service for many applied psychologists. Their proprietary nature is fundamental to the perpetuation of these royalty streams and to the privileged status of trained psychologists. For industrial and clinical settings, copyright-protected commercial measures offer clear benefits.

However, the needs of primary researchers often differ from those of commercial test users. These differences relate to issues of score interpretation, test content and administrative flexibility. In the case of score interpretation, researchers are considerably less concerned about the nature and quality of interpretative feedback. Unlike test-takers in selection and clinical settings, research participants are typically motivated by monetary rewards, course credit or, perhaps, a casual desire for informal feedback about their performance. This does not imply that researchers are less interested in quality norming data – it is often critical for evaluating the degree to which a sample is representative of a broader population. It simply means that, while many commercial testing companies have attempted to differentiate their products by providing materials for individual score interpretation, these materials have relatively little value for administration in research contexts.

The motivation among commercial testing companies to provide useful interpretative feedback is directly related to test content however, and the nature of test content is of critical importance for intelligence researchers. The typical rationale for cognitive ability assessment in research settings is to evaluate the relationship between constructs and a broad range of other attributes. As such, the variety and depth of a test's content are very meaningful criteria for intelligence researchers – ones which are somewhat incompatible with the provision of meaningful interpretative feedback for each type of content. In other words, the ideal circumstance for many researchers would include the ability to choose from a variety of broadly-assessed cognitive ability constructs (or perhaps to choose a single measure which includes the assessment of a broad variety of constructs). While this ideal can sometimes be achieved through the administration of multiple commercial measures, this is rarely practical due to issues of cost and/or a lack of administrative flexibility.

The cost of administering commercial tests in research settings varies considerably across measures. While published rates are typically high, many companies allow for the qualified use of their copyright-protected materials at reduced rates or free-of-charge in research settings (e.g., the ETS Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976)). Variability in administration and scoring procedures is similarly high across measures. A small number of extant tests allow for brief, electronic assessment with automated scoring conducted within the framework of proprietary software, though none of these measures allow for customization of test content. The most commonly-used batteries are more arduous to administer, requiring one-to-one administration for over an hour followed by an additional 10 to

20 minutes for scoring (Camara et al., 2000). All too often, the result of the combination of challenges posed by these constraints is the omission of cognitive ability assessment in psychological research.

Several authors have suggested that the pace of scientific progress is diminished by reliance on proprietary measures (Gambardella and Hall, 2006; Goldberg, 1999; Liao et al., 2008). While it is difficult to evaluate this claim empirically in the context of intelligence research, the circumstances surrounding development of the International Personality Item Pool ("IPIP") (Goldberg, 1999; Goldberg et al., 2006) provide a useful analogy. Prior to the development of the IPIP, personality researchers were forced to choose between validated but restrictive proprietary measures and a disorganized collection of narrow-bandwidth public-domain scales (these having been developed by researchers who were either unwilling to deal with copyright issues or whose needs were not met by the content of proprietary options). In the decade ending in 2012, at least 500 journal articles and book chapters using IPIP measures were published (Goldberg, 2012).

In fact, most of the arguments set forth in Goldberg's (1999) proposal for public-domain measures are directly applicable here. His primary point was that unrestricted use of public-domain instruments would make it less costly and difficult for researchers to administer scales which are flexible and widely-used. Secondary benefits would include a collaborative medium through which researchers could contribute to test development, refinement, and validation. The research community as a whole would benefit from an improved means of empirically comparing hypotheses across many diverse criteria.

Critics of the IPIP proposal expressed concern that a lack of copyright protection would impair the validity of personality measures (Goldberg et al., 2006). This argument would seem even more germane for tests of cognitive ability given the "maximal performance/typical behavior" distinction between intelligence and personality measures. The widely-shared presumption is that copyright restrictions on proprietary tests maintain validity by enhancing test security. Testing materials are, in theory, only disseminated to authorized users who have purchased licensed access and further dissemination is discouraged by the enforcement of intellectual property laws. Unfortunately, it is difficult to ascertain the extent to which test validity would be compromised in the general population without these safeguards. Concerns about disclosure have been called into question with several prominent standardized tests (Field, 2012). There is also debate about the efficacy of intellectual property laws for protection against the unauthorized distribution of testing materials via the internet (Field, 2012; Kaufmann, 2009; McCaffrey and Lynch, 2009). Further evaluation of the relationship between copyright-protection and test validity seems warranted by these concerns, particularly for research applications where individual outcomes are less consequential.

Fortunately, copyright protection is not a prerequisite for test validity. Modern item-generation techniques (Arendasy et al., 2006; Dennis et al., 2002) present an alternate strategy that is less dependent on test security. Automatic item-generation makes use of algorithms which dictate the parameters of new items with predictable difficulty and in many alternate forms. These techniques allow for the creation of item types where the universe of *possible* items is very large. This, in turn, reduces the threat to validity that results from item disclosure. It can even be used to enhance test validity under administration paradigms that expose participants to sample items prior to testing and use alternate forms during assessment as this methodology reduces the effects of differential test familiarity across participants.

While automatic item-generation techniques represent the optimal method for developing public-domain cognitive ability items, this approach is often considerably more complicated than traditional development methods and it may be some time before a sizable number of automatically-generated item types is available for use in the public domain. For item types developed by traditional means, the maintenance of test validity depends on implementation of the more practical protocols used by commercial measures (i.e., those which do not invoke the credible threat of legal action). A public domain resource should set forth clear expectations for researchers regarding appropriate and ethical usage and make use of "warnings for nonprofessionals" (Goldberg et al., 2006). Sample test items should be made easily available to the general public to further discourage wholesale distribution of testing materials. Given the current barriers to enforcement for intellectual property holders, these steps are arguably commensurate with protocols in place for copyright-protected commercial measures.

To the extent that traditional and automatic item-generation methods maintain adequate validity, there are many applications in which a non-proprietary measure would be useful. The most demanding of these applications would involve distributed, un-proctored assessments *in situ*, presumably conducted via online administration. Validity concerns would be most acute in these situations as there would be no safeguards against the use of external resources, including those available on the internet.

The remainder of this paper is dedicated to the evaluation of a public-domain measure developed for use under precisely these circumstances. This measure, the International Cognitive Ability Resource ("ICAR"), has been developed in stages over several years and further development is on-going. The first four item types (described below) were initially designed to provide an estimation of general cognitive ability for participants completing personality surveys at SAPA-Project.org, previously test.personality-project.org.

The primary goals when developing these initial item types were to: (1) briefly assess a small number of cognitive ability domains which were relatively distinct from one another (though considerable overlap between scores on the various types was anticipated); (2) avoid the use of "timed" items in light of potential technical issues resulting from telemetric assessment (Wilt et al., 2011); and (3) avoid item content that could be readily referenced elsewhere given the intended use of un-proctored online administrations. The studies described below were conducted to evaluate the degree to which these goals of item development were achieved.

The first study evaluated the item characteristics, reliability and structural properties of a 60-item ICAR measure. The second study evaluated the validity of the ICAR items when administered online in the context of self-reported achievement test scores and university majors. The third study evaluated the construct validity of the ICAR items when administered offline, using a brief commercial measure of cognitive ability.

3. Study 1

We investigated the structural properties of the initial version of the International Cognitive Ability Resource based on internet administration to a large international sample. This investigation was based on 60 items representing four item types developed in various stages since 2006 (and does not include deprecated items or item types currently under development). We hypothesized that the factor structure would demonstrate four distinct but highly correlated factors, with each type of item represented by a separate factor. This implied that, while individual items might demonstrate moderate or strong cross-loadings, the primary loadings would be consistent among items of each type.

3.1. Method

3.1.1. Participants

Participants were 96,958 individuals (66% female) from 199 countries who completed an online survey at SAPA-project.org (previously test.personality-project.org) between August 18, 2010 and May 20, 2013 in exchange for customized feedback about their personalities. All data were self-reported. The mean self-reported age was 26 years (sd = 10.6, median = 22) with a range from 14 to 90 years. Educational attainment levels for the participants are given in Table 1. Most participants were current university or secondary school students, although a wide range of educational attainment levels were represented. Among the 75,740 participants from the United States (78.1%), 67.5% identified themselves as White/Caucasian, 10.3% as African-American, 8.5% as Hispanic-American, 4.8% as Asian-American, 1.1% as

Native-American, and 6.3% as multi-ethnic (the remaining 1.5% did not specify). Participants from outside the United States were not prompted for information regarding race/ethnicity.

3.1.2. Measures

Four item types from the International Cognitive Ability Resource were administered, including: 9 Letter and Number Series items, 11 Matrix Reasoning items, 16 Verbal Reasoning items and 24 Three-Dimensional Rotation items. A 16 item subset of the measure, hereafter referred to as the *ICAR Sample Test*, is included as Appendix A in the Supplemental Materials. [1] Letter and Number Series items prompt participants with short digit or letter sequences and ask them to identify the next position in the sequence from among six choices. Matrix Reasoning items contain stimuli that are similar to those used in Raven's Progressive Matrices. The stimuli are 3x3 arrays of geometric shapes with one of the nine shapes missing. Participants are instructed to identify which of six geometric shapes presented as response choices will best complete the stimuli. The Verbal Reasoning items include a variety of logic, vocabulary and general knowledge questions. The Three-Dimensional Rotation items present participants with cube renderings and ask participants to identify which of the response choices is a possible rotation of the target stimuli. None of the items were timed in these admin-

[1] In addition to the sample items available in Appendix A, the remaining ICAR items can be accessed through ICAR-Project.org. A sample data set based on the items listed in Appendix A is also available ('iqitems') through the *psych* package (Revelle, 2013) in the R computing environment (R Core Team, 2013).

istrations as untimed administration was expected to provide more stringent and conservative evaluation of the items' utility when given online (there are no specific reasons precluding timed administrations of the ICAR items, whether online or offline).

Participants were administered 12 to 16 item subsets of the 60 ICAR items using the Synthetic Aperture Personality Assessment ("SAPA") technique (Revelle et al., 2010), a variant of matrix sampling procedures discussed by Lord (1955). The number of items administered to each participant varied over the course of the sampling period and was independent of participant characteristics. The number of administrations for each item varied considerably (median = 21,764) as did the number of pairwise administrations between any two items in the set (median = 2,610). This variability reflected the introduction of newly developed items over time and the fact that item sets include unequal numbers of items. The minimum number of pairwise administrations among items (422) provided sufficiently high stability in the covariance matrix for the structural analyses described below (Kenny, 2012).

3.1.3. Analyses

Internal consistency measures were assessed by using the Pearson correlations between ICAR items to calculate α, ω_h, and ω_{total} reliability coefficients (Revelle, 2013; Revelle and Zinbarg, 2009; Zinbarg et al., 2005). The use of tetrachoric correlations for reliability analyses is discouraged on the grounds that it typically over-estimates both alpha and omega (Revelle and Condon, 2012).

Two latent variable exploratory factor analyses ("EFA") were conducted to evaluate the structure of the ICAR items. The first of these included all

60 items (9 Letter and Number Series items, 11 Matrix Reasoning items, 16 Verbal Reasoning items and 24 Three-Dimensional Rotation items). A second EFA was required to address questions regarding the structural impact of including disproportionate numbers of items by type. This was done by using only the subset of participants (n = 4,574) who were administered the 16 item *ICAR Sample Test*. This subset included four items each from the four ICAR item types. These items were selected as a representative set on the basis of their difficulty relative to the full set of 60 items and their factor loadings relative to other items of the same type. Note that the factor analysis of this 16 item subset was not independent from that conducted on the full 60 item set. EFA results were then used to evaluate the omega hierarchical general factor saturation (Revelle and Zinbarg, 2009; Zinbarg et al., 2006) of the 16 item *ICAR Sample Test*.

Both of these exploratory factor analyses were based on the Pearson correlations between scored responses using Ordinary Least Squares ("OLS") regression models with oblique rotation (Revelle, 2013). The factoring method used here minimizes the χ^2 value rather than minimizing the sum of the squared residual values (as is done by default with most statistical software). Note that in cases where the number of administrations is consistent across items, as with the 16 item *ICAR Sample Test*, these methods are identical. The methods differ in cases where the number of pairwise administrations between items varies because the squared residuals are weighted by sample size rather than assumed to be equivalent across variables. Goodness-of-fit was evaluated using the Root Mean Square of the Residual, the Root Mean Squared Error of Approximation (Hu and Bentler, 1999), and the Tucker

Lewis Index of factoring reliability (Kenny, 2012; Tucker and Lewis, 1973).

Analyses based on two-parameter Item Response Theory (Baker, 1985; Embretson, 1996; Revelle, 2013) were used to evaluate the unidimensional relationships between items on several levels, including (1) all 60 items, (2) each of the four item types independently, and (3) for the 16 item *ICAR Sample Test*. In these cases, the tetrachoric correlations between items were used. These procedures allow for estimation of the correlations between items as if they had been measured continuously (Uebersax, 2000).

3.2. Results

Descriptive statistics for all 60 ICAR items are given in Table 2. Mean values indicate the proportion of participants who provided the correct response for an item relative to the total number of participants who were administered that item. The Three-Dimensional Rotation items had the lowest proportion of correct responses (m = 0.19, sd = 0.08), followed by Matrix Reasoning (m = 0.52, sd = 0.15), then Letter and Number Series (m = 0.59, sd = 0.13), and Verbal Reasoning (m = 0.64, sd = 0.22). Internal consistencies for the ICAR item types are given in Table 3. These values are based on the composite correlations between items as individual participants completed only a subset of the items (as is typical when using SAPA sampling procedures).

Results from the first exploratory factor analysis using all 60 items suggested factor solutions of three to five factors based on inspection of the scree plots in Figure 1. The fit statistics were similar for each of these solutions. The four factor model was slightly superior in fit (RMSEA = 0.058, RMSR = 0.05) and reliability (TLI = 0.71) to the three factor model (RMSEA =

0.059, RMSR = 0.05, TLI = 0.7) and was slightly inferior to the five factor model (RMSEA = 0.055, RMSR = 0.05, TLI = 0.73). Factor loadings and the correlations between factors for each of these solutions are included in the supplementary materials (see Supplementary Tables 1 to 6).

The second EFA, based on a balanced number of items by type, demonstrated very good fit for the four-factor solution (RMSEA = 0.014, RMSR = 0.01, TLI = 0.99). Factor loadings by item for the four-factor solution are shown in Table 4. Each of the item types was represented by a different factor and the cross-loadings were small. Correlations between factors (Table 5) ranged from 0.41 to 0.70.

General factor saturation for the 16 item *ICAR Sample Test* is depicted in Figures 2 and 3. Figure 2 shows the primary factor loadings for each item consistent with the values presented in Table 4 and also shows the general factor loading for each of the second-order factors. Figure 3 shows the general factor loading for each item and the residual loading of each item to its primary second-order factor after removing the general factor.

The results of IRT analyses for the 16 item *ICAR Sample Test* are presented in Table 6 as well as Figures 4 and 5. Table 6 provides item information across levels of the latent trait and summary information for the test as a whole. The item information functions are depicted graphically in Figure 4. Figure 5 depicts the test information function for the *ICAR Sample Test* as well as reliability in the vertical axis on the right (reliability in this context is calculated as one minus the reciprocal of the test information). The results of IRT analyses for the full 60 item set and for each of the item types independently are available in the supplementary materials (Supplementary

Tables 7 to 11). The pattern of results was similar to those for the *ICAR Sample Test* in terms of the relationships between item types and the spread of item difficulties across levels of the latent trait, though the reliability was higher for the full 60 item set across the range of difficulties (Supplementary Figure 1).

3.3. Discussion

A key finding from Study 1 relates to the broad range of means and standard deviations for the ICAR items as these values demonstrated that the un-proctored and untimed administration of cognitive ability items online does not lead to uniformly high scores with insufficient variance. To the contrary, all of the Three-Dimensional Rotation items and more than half of all 60 items were answered incorrectly more often than correctly and the weighted mean for all items was only 0.53. This point was further supported by the IRT analyses in that the item information functions demonstrate a relatively wide range of item difficulties.

Internal consistency was good for the Three-Dimensional Rotation item type, adequate for the Letter and Number Series and the Verbal Reasoning item types, and marginally adequate for the Matrix Reasoning item type. This suggests that the 11 Matrix Reasoning items were not uniformly measuring a singular latent construct whereas performance on the Three-Dimensional Rotation items was highly consistent. For the composites based on both 16 and 60 items however, internal consistencies were adequate (α=0.81; ω_{total}=0.83) and good (α=0.93; ω_{total}=0.94), respectively. While higher reliabilities reflect the greater number of items in the ICAR60, it should be noted that the general factor saturation was slightly higher for the shorter 16-item measure (ICAR16 ω_h=0.66; ICAR60 ω_h=0.61). When considered as a function of test information, reliability was generally adequate across a wide range of latent trait levels, and particularly good within approximately ±1.5 standardized units from the mean item difficulty. All of the factor analyses demonstrated evidence of both a positive manifold among items and high general factor saturation for each of the item types. In the four factor solution for the 16 item scale, the Verbal Reasoning and the Letter and Number Series factors showed particularly high *'g'* loadings (0.8).

4. Study 2

Following the evidence for reliable variability in ICAR scores in Study 1, it was the goal of Study 2 to evaluate the validity of these scores when using the same administration procedures. While online administration protocols precluded validation against copyrighted commercial measures, it was possible to evaluate the extent to which ICAR scores correlated with (1) self-reported achievement test scores and (2) published rank orderings of mean scores by university major. In the latter case, ICAR scores were expected to demonstrate group discriminant validity by correlating highly with the rank orderings of mean scores by university major as previously described by the Educational Testing Service (Educational Testing Service, 2010) and the College Board (College Board, 2012).

In the former case, ICAR scores were expected to reflect a similar relationship with achievement test scores as extant measures of cognitive ability. Using data from the National Longitudinal Study of Youth 1979, Frey and Detterman (2004) reported simple correlations between the SAT and the

Armed Services Vocational Aptitude Battery (r = 0.82, n = 917) and several additional IQ measures (rs = 0.53 - 0.82) with smaller samples (ns = 15 - 79). In a follow-up study with a university sample, Frey and Detterman (2004) evaluated the correlation between combined SAT scores and Raven's Progressive Matrices scores, finding an uncorrected correlation of 0.48 ($p < .001$) and a correlation after correcting for restriction of range of 0.72. Similar analyses with ACT composite scores (Koenig et al., 2008) showed a correlation of 0.77 ($p < .001$) with the ASVAB, an uncorrected correlation with the Raven's Advanced Progressive Matrices of 0.61 ($p < .001$), and a correlation corrected for range restriction with the Raven's APM of 0.75.

Given the breadth and duration of assessment for the ASVAB, the SAT and the ACT, positive correlations of a lesser magnitude were expected between the ICAR scores and the achievement tests than were previously reported with the ASVAB. Correlations between the Raven's APM and the achievement test scores were expected to be more similar to the correlations between the achievement test scores and the ICAR scores, though it was not possible to estimate the extent to which the correlations would be affected by methodological differences (i.e., the un-proctored online administration of relatively few ICAR items and the use of self-reported, rather than independently verified, achievement test scores as described in the Methods section below).

4.1. Method

4.1.1. Participants

The 34,229 participants in Study 2 were a subset of those used for Study 1, chosen on the basis of age and level of educational attainment. Participants were 18 to 22 years old (m = 19.9, $s.d.$ = 1.3, median = 20). Approximately 91% of participants had begun but not yet attained an undergraduate degree; the remaining 9% had attained an undergraduate degree. Among the 26,911 participants from the United States, 67.1% identified themselves as White/Caucasian, 9.8% as Hispanic-American, 8.4% as African-American, 6.0% as Asian-American, 1.0% as Native-American, and 6.3% as multi-ethnic (the remaining 1.5% did not specify).

4.1.2. Measures

Both the sampling method and the ICAR items used in Study 2 were identical to the procedures described in Study 1, though the total item administrations (median = 7,659) and pairwise administrations (median = 906) were notably fewer given that the participants in Study 2 were a sub-sample of those in Study 1. Study 2 also used self-report data for three additional variables collected through SAPA-project.org: (1) participants' academic major on the university level, (2) their achievement test scores, and (3) participants' scale scores based on randomly administered items from the Intellect scale of the "100-Item Set of IPIP Big-Five Factor Markers" (Goldberg, 2012). For university major, participants were allowed to select only one option from 147 choices, including "undecided" (n = 3,460) and several categories of "other" based on academic disciplines. For the achievement test scores, participants were given the option of reporting 0, 1, or multiple types of scores, including: SAT Critical Reading (n = 7,404); SAT Mathematics (n = 7,453); and the ACT (n = 12,254). Intellect scale scores were calculated using IRT procedures, assuming unidimensionality for the Intellect items only (items assessing Openness were omitted). Based on composites of the Pearson correlations between items without imputation of missing values, the Intellect scale had an α of 0.74, an ω_h of 0.60, and an ω_{total} of 0.80. The median number of pairwise administrations for these items was 4,475.

4.1.3. Analyses

Two distinct methods were used to calculate the correlations between the achievement test scores and the ICAR scores in order to evaluate the effects of two different corrections. The first method used ICAR scale scores based on composites of the tetrachoric correlations between ICAR items (composites are used because each participant was administered 16 or fewer items). The correlations between these scale scores and the achievement test scores were then corrected for reliability. The α reliability coefficients reported in Study 1 were used for the ICAR scores. For the achievement test scores, the need to correct for reliability was necessitated by the use of self-reported scores. Several researchers have demonstrated the reduced reliability of self-reported scores in relation to official test records (Cassady, 2001; Cole and Gonyea, 2009; Kuncel et al., 2005; Mayer et al., 2006), citing participants' desire to misrepresent their performance and/or memory errors as the most likely causes. Despite these concerns, the reported correlations between self-reported and actual scores suggest that the rank-ordering of scores is maintained, regardless of the magnitude of differences (Cole and Gonyea, 2009; Kuncel et al., 2005; Mayer et al., 2006). Reported correlations between self-reported and actual scores have ranged from 0.74 to 0.86 for the SAT - Critical Reading section, 0.82 to 0.88 for the SAT - Mathematics, and 0.82 to 0.89 for the SAT - Combined (Cole and Gonyea, 2009; Kuncel et al., 2005; Mayer et al., 2006). Higher correlations were found by Cole and Gonyea

(2009) for the ACT Composite (0.95). The Study 2 sample approximated the samples on which these reported correlations were based in that (1) participants were reminded about the anonymity of their responses and (2) the age range of participants was limited to 18 to 22 years. The weighted mean values from these findings (SAT-CR = 0.86; SAT-M = 0.88; SAT-Combined = 0.88; ACT = 0.95) were used as reliability coefficients for the achievement test scores when correcting correlations between the achievement tests and other measures (ICAR scores and the IPIP-100 Intellect scores).

The second method for calculating correlations between ICAR scores and achievement test scores used IRT-based (2PL) scoring (Revelle, 2013). Scale scores for each item type and the full test were calculated for each participant, and these scale scores were then correlated with the achievement test scores. In this case, corrections were made to address the potential for an incidental selection effect due to optional reporting of achievement test scores (Cassady, 2001; Frucot and Cook, 1994). 52.5% of participants in Study 2 did not report any achievement test scores; 10.1% reported scores for all three (SAT - CR, SAT - M, and ACT). These circumstances would result in an incidental selection effect if the correlations between self-reported achievement test scores and the ICAR measures were affected by the influence of a third variable on one or both measures (Sackett and Yang, 2000). The so-called "third" variable in this study likely represented a composite of latent factors which are neither ergodic nor quantifiable but which resulted in group differences between those who reported their scores and those who did not. If the magnitude of differences in achievement test scores between groups were non-trivial, the effect on the overall correlations would also be non-trivial given the proportion of participants not reporting. The need for correction procedures in this circumstance was elaborated by both Pearson (1903) and Thorndike (1949), though the methods employed here were developed in the econometrics literature and are infrequently used by psychologists (Sackett and Yang, 2000). Clark and Houle (2012) and Cuddeback et al. (2004) provide useful illustrations of these procedures. The two-step method of the "Heckman correction" (Greene, 2008; Heckman, 1976, 1979; Toomet and Henningsen, 2008) was used to evaluate and correct for selection effects where warranted using IPIP-100 Intellect scores.

In addition to these analyses of the relationship between ICAR scores and achievement test scores, the Study 2 sample was used to evaluate the correlations between the ICAR items and the published rank orderings of mean scores by university major. This was done using IRT-based ICAR scores when grouped by academic major on the university level. These were evaluated relative to similar data sets published by the Educational Testing Service (Educational Testing Service, 2010) and the College Board (College Board, 2012) for the GRE and SAT, respectively. GRE scores were based on group means for 287 "intended graduate major" choices offered to fourth-year university students and non-enrolled graduates who took the GRE between July 1, 2005 and June 30, 2008 (N = 569,000). These 287 groups were consolidated with weighting for sample size in order to match the 147 university major choices offered with the ICAR. Of these 147 majors, only the 91 with $n > 20$ were used. SAT scores were based on group means for 38 "intended college major" choices offered to college-bound seniors in the high school graduating class of 2012 (N = 1,411,595). In this case, the 147 university major choices offered with the ICAR were consolidated to match 29 of the choices offered with the SAT. The 9 incompatible major choices collectively represented only 1.3% of the SAT test-takers. The omitted majors were: Construction Trades; Mechanic and Repair Technologies/Technician; Military Technologies and Applied Sciences; Multi/Interdisciplinary Studies; Precision Production; Security and Protective Services; Theology and Religious Vocations; Other; and Undecided.

4.2. Results

Descriptive statistics for the self-reported achievement test scores are shown in Table 7. Correlations between self-reported achievement test scores and ICAR scale scores calculated using composites of the tetrachoric correlations are shown in Table 8, with uncorrected correlations shown below the diagonal and the correlations corrected for reliability shown above the diagonal. Reliabilities for each measure are given on the diagonal. Correlations between composites which were not independent have been omitted. Corrected correlations between the achievement test scores and both the 16 and 60 item ICAR composites ranged from 0.52 - 0.59 (ses ≤ 0.016).[2]

Table 9 presents the correlations between the self-reported achievement test scores and the IRT-based ICAR scores, with the uncorrected correlations below the diagonal and the correlations corrected for incidental selection

[2]The standard error of the composite scores are a function of both the number of items and the number of participants who took each pair of items (Revelle and Brown, 2013). Estimates of the standard errors can be identified through the use of bootstrapping procedures to derive estimates of the confidence intervals of the correlations (Revelle, 2013). In this case, the confidence intervals were estimated based on 100 sampling iterations.

effects above the diagonal. Correlations between non-independent scores were omitted. Scores for the ICAR measures were based on a mean of 2 to 4 responses for each of the item types (mean number of LN items administered $= 3.2$, $sd = 1.3$; MR items $m = 2.8$, $sd = 1.1$; R3D items $m = 2.0$, $sd = 1.5$; VR items $m = 4.3$, $sd = 2.2$) and 12 to 16 items for the ICAR60 scores ($m = 12.4$, $sd = 3.8$). Corrected correlations between the achievement test scores and ICAR60 ranged from 0.44 to 0.47 (ses ≤ 0.016).

Tables 10 and 11 contain group-level correlations using mean scores for university major. Table 10 shows the correlations between the published norms for the SAT, the mean self-reported SAT scores for each major in the Study 2 sample, and the mean IRT-based ICAR scores for each major in the Study 2 sample. The correlation between mean ICAR scores by major and mean combined SAT scores by major in the published norms was 0.75 ($se =$ 0.147). Table 11 shows the correlations between the published norms for the GRE by major and the IRT-based ICAR scores for the corresponding majors in the Study 2 sample (self-reported GRE scores were not collected). The correlation between mean ICAR scores by major and mean combined GRE scores by major in the published norms was 0.86 ($se = 0.092$).

4.3. Discussion

After correcting for the "reliability" of self-reported scores, the 16 item *ICAR Sample Test* correlated 0.59 with combined SAT scores and 0.52 with the ACT composite. Correlations based on the IRT-based ICAR scores were lower though these scores were calculated using even fewer items; correlations were 0.47 and 0.44 with combined SAT scores and ACT composite scores respectively based on an average of 12.4 ICAR60 items answered per participant. As expected, these correlations were smaller than those reported for longer cognitive ability measures such as the ASVAB and the Raven's APM (Frey and Detterman, 2004; Koenig et al., 2008).

The ICAR items demonstrated strong group discriminant validity on the basis of university majors. This indicates that the rank ordering of mean ICAR scores by major is strongly correlated with the rank ordering of mean SAT scores and mean GRE scores. Consistent with the individual-level correlations, the group-level correlations were higher between the ICAR subtests and the mathematics subtests of the SAT and the GRE relative to the verbal subtests.

5. Study 3

The goal of the third study was to evaluate the construct validity of the ICAR items against a commercial measure of cognitive ability. Due to the copyrights associated with commercial measures, these analyses were based on administration to an offline sample of university students rather than an online administration.

5.1. Method

5.1.1. Participants

Participants in Study 3 were 137 college students (76 female) enrolled at a selective private university in the midwestern United States. Students participated in exchange for credit in an introductory psychology course. The mean age of participants in this sample was 19.7 years (sd = 1.2, median = 20) with a range from 17 to 25 years. Within the sample, 67.2% reported being first-year students, 14.6% second-year students, 8.0% third-year students and the remaining 10.2% were in their fourth year or beyond. With regards to ethnicity, 56.2% identified themselves as White/Caucasian, 26.3% as Asian-American, 4.4% as African-American, 4.4% as Hispanic-American, and 7.3% as multi-ethnic (the remaining 1.5% did not specify).

5.1.2. Measures

Participants in the university sample were administered the 16 item *ICAR Sample Test*. The presentation order of these 16 items was randomized across participants. Participants were also administered the *Shipley-2*, which is a 2009 revision and restandardization of the *Shipley Institute of Living Scale* (Shipley et al., 2009, 2010). The *Shipley-2* is a brief measure of cognitive functioning and impairment that most participants completed in 15 to 25 minutes. While the *Shipley-2* is a timed test, the majority of participants stopped working before using all of the allotted time. The *Shipley-2* has two administration options. Composite A ($n = 69$) includes a vocabulary scale designed to assess crystallized skills and an abstraction scale designed to assess fluid reasoning skills (Shipley et al., 2009). Composite B ($n = 68$) includes the same vocabulary scale and a spatial measure of fluid reasoning called the "Block Patterns" scale (Shipley et al., 2009). All three scales included several items of low difficulty with little or no variance in this sample. After removal of items without variance, internal consistencies were low for the Abstraction scale (10 of 25 items removed, $\alpha = 0.37$; $\omega_{total} = 0.51$) and the Vocabulary scale (7 of 40 items removed, $\alpha = 0.61$; $\omega_{total} = 0.66$). The Block Patterns scale had fewer items without variance (3 of 26) and adequate consistency ($\alpha = 0.83$, $\omega_{total} = 0.88$). Internal consistencies were calculated using Pearson correlations between items.

5.1.3. Analyses

Correlations were evaluated between scores on the *ICAR Sample Test* and a brief commercial measure of cognitive ability, the *Shipley-2*. Two types of corrections were relevant to these correlations; one for the restriction of range among scores and a second for reliability. The prospect of range restriction was expected on the grounds that participants in the sample were students at a highly selective university. The presence of restricted range was evaluated by looking for reduced variance in the sample relative to populations with similar characteristics. In this case, the university sample was evaluated relative to the online sample. Where present, the appropriate method for correcting this type of range restriction uses the following equation (case 2c from Sackett and Yang, 2000) (Bryant and Gokhale, 1972; Alexander, 1990):

$$\hat{\rho}_{xy} = r_{xy}(s_x/S_x)(s_y/S_y) \pm \sqrt{[1-(s_x/S_x)^2][1-(s_y/S_y)^2]} \tag{1}$$

where s_x and s_y are the standard deviations in the restricted sample, S_x and S_y are the standard deviations in the unrestricted sample and the $\pm$ sign is conditional on the direction of the relationship between the selection effect and each of the variables, x and y. When correcting for reliability, the published reliabilities (Shipley et al., 2010) were used for each of the *Shipley-2* composites (0.925 for Composite A and 0.93 for Composite B) instead of the reliabilities within the sample due to the large number of items with little or no variance.

5.2. Results

The need to correct for restriction of range was indicated by lower standard deviations of scores on all of the subtests and composites for the *Shipley-2* and the *ICAR Sample Test*. Table 12 shows the standard deviation of scores for the participants in Study 3 (the "restricted" sample) and the reference scores (the "unrestricted" samples).

Correlations between the ICAR scores and *Shipley-2* scores are given in Table 13, including the uncorrected correlations, the correlations corrected for range restriction and the correlations corrected for reliability and range restriction. The range and reliability corrected correlations between the *ICAR Sample Test* and the *Shipley-2* composites were nearly identical at 0.81 and 0.82 ($se = 0.10$).

5.3. Discussion

Correlations between the ICAR scores and the *Shipley-2* were comparable to those between the *Shipley-2* and other measures of cognitive ability. The correlations after correcting for reliability and restricted range between the 16 item *ICAR Sample Test* and *Shipley-2* composite A and B were 0.82 and 0.81, respectively. Correlations between *Shipley-2* composite A and B were 0.64 and 0.60 with the *Wonderlic Personnel Test*, 0.77 and 0.72 with the Full-Scale IQ scores for the *Wechsler Abbreviated Scale of Intelligence* in an adult sample, and 0.86 and 0.85 with the Full-Scale IQ scores for the *Wechsler Adult Intelligence Scale* (Shipley et al., 2010).

6. General Discussion

Reliability and validity data from these studies suggest that a public-domain measure of cognitive ability is a viable option. More specifically, they demonstrate that brief, un-proctored, and untimed administrations of items from the International Cognitive Ability Resource are moderately-to-strongly correlated with measures of cognitive ability and achievement. While this method of administration is inherently less precise and exhaustive than many traditional assessment methods, it offers many benefits. Online assessment allows for test administration at any time of day, in any geographic location, and over any type of internet-enabled electronic device. These administrations can be conducted either with or without direct interaction with the research team. Measures constructed with public-domain item types like those described here can be easily customized for test length and content as needed to match the research topic under evaluation. All of this can be accomplished without the cost, licensing, training, and software needed to administer the various types of copyright-protected commercial measures.

These data also suggest that there are many ways in which the ICAR can be improved. With regard to the existing item types, more - and more difficult - items are needed for all of the item types except perhaps the Three-Dimensional Rotation items. While the development of additional Letter and Number Series items can be accomplished formulaically, item development procedures for the Verbal Reasoning items is complicated by the need for items to be resistant to basic internet word searches. The Matrix Reasoning items require further structural analyses before further item development as these items demonstrated less unidimensionality than the other three item types. This may be appropriate if they are to be used as a measure of general cognitive ability, but it remains important to identify the ways in which these items assess subtly different constructs. This last point relates to the additional need for analyses of differential item functioning for all of the item types and the test as a whole.

The inclusion of many more item types in the ICAR is also needed as is more extensive validation of new and existing item types. The most useful additions in the near term would include item types which assess constructs distinct from the four item types described here. Several such item types are in various stages of development and piloting by the authors and their collaborators. These item types should be augmented with extant, public-domain item types when feasible.

7. Conclusion

Public-domain measures of cognitive ability have considerable potential. We propose that the International Cognitive Ability Resource provides a viable foundation for collaborators who are interested in contributing extant or newly-developed public-domain tools. To the extent that these tools are well-suited for online administration, they will be particularly useful for large-scale cognitive ability assessment and/or use in research contexts beyond the confines of traditional testing environments. As more item types become available, the concurrent administration of ICAR item types will become increasingly valuable for researchers studying the structure of cognitive abilities on both the broad, higher-order levels (e.g., spatial and verbal abilities) as well as the relatively narrow (e.g., more closely related abilities such as two- and three-dimensional rotation). The extent to which a public-domain resource like the ICAR fulfills this potential ultimately depends on the researchers for whom it offers the highest utility. We entreat these potential users to consider contributing to its on-going development, improvement, validation and maintenance.

References

Alexander, R. A. (1990). Correction formulas for correlations restricted by selection on an unmeasured variable. *Journal of Educational Measurement*, 27(2):187–189.

Arendasy, M., Sommer, M., Gittler, G., and Hergovich, A. (2006). Automatic generation of quantitative reasoning items. *Journal of Individual Differences*, 27(1):2–14.

Baker, F. B. (1985). *The basics of item response theory*. Heinemann Educational Books, Portsmouth, NH.

Bryant, N. D. and Gokhale, S. (1972). Correcting correlations for restrictions in range due to selection on an unmeasured variable. *Educational and Psychological Measurement*, 32(2):305–310.

Camara, W. J., Nathan, J. S., and Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, 31(2):141–154.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press, Cambridge, UK.

Cassady, J. C. (2001). Self-reported GPA and SAT: A methodological note. *Practical Assessment, Research & Evaluation*, 7(12).

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin*, 40(3):153–193.

Clark, S. J. and Houle, B. (2012). Evaluation of Heckman selection model method for correcting estimates of HIV prevalence from sample surveys. *Center for Statistics and the Social Sciences*, Working Paper no. 120:1–18.

Cole, J. S. and 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: The College Board. Retrieved September 13, 2013, from http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf.

Cuddeback, G., Wilson, E., Orme, J. G., and Combs-Orme, T. (2004). Detecting and statistically correcting sample selection bias. *Journal of Social Service Research*, 30(3):19–33.

Dennis, I., Handley, S., Bradon, P., Evans, J., and Newstead, S. (2002). Approaches to modeling item-generative tests. In Irvine, S. H. and Kyllonen, P. C., editors, *Item generation for test development*, pages 53–71. Lawrence Erlbaum Associates, Mahwah, New Jersey.

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

Ekstrom, R. B., French, J. W., Harman, H. H., and Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Educational Testing Service, Princeton, NJ.

Eliot, J. and Smith, I. M. (1983). *An International Directory of Spatial Tests*. NFER-NELSON Publishing Company Ltd., Great Britain.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8(4):341–349.

Field, T. G. (2012). *Standardized tests: Recouping development costs and preserving integrity.* Retrieved September 13, 2013, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1989584.

Frey, M. C. and Detterman, D. K. (2004). Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability. *Psychological Science*, 15(6):373–378.

Frucot, V. and Cook, G. (1994). Further research on the accuracy of students' self-reported grade point averages, SAT scores, and course grades. *Perceptual and Motor Skills*, 79(2):743–746.

Gambardella, A. and Hall, B. H. (2006). Proprietary versus public domain licensing of software and research products. *Research Policy*, 35(6):875–892.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several Five-Factor Models. In Mervielde, I., Deary, I., De Fruyt, F., and Ostendorf, F., editors, *Personality Psychology in Europe*, pages 1–7. Tilburg University Press, Tilburg, The Netherlands.

Goldberg, L. R. (2012). *International Personality Item Pool: A scientific collaboratory for the development of advanced measures of personality* *traits and other individual differences.* Retrieved November 16, 2012, from http://ipip.ori.org/.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., and Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1):84–96.

Goldstein, G. and Beers, S. R., editors (2004). *Comprehensive Handbook of Psychological Assessment, Volume 1: Intellectual and Neuropsychological Assessment.* John Wiley & Sons, Inc., Hoboken, NJ.

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

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 Berg, S. V., editor, *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 475–492. NBER, Cambridge, MA.

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

Hu, L. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1):1–55.

Kaufmann, P. (2009). Protecting raw data and psychological tests from wrongful disclosure: A primer on the law and other persuasive strategies. *The Clinical Neuropsychologist*, 23(7):1130–1159.

Kenny, D. A. (2012). *Measuring model fit.* Retrieved November 7, 2012, from http://www.davidakenny.net/cm/fit.htm.

Koenig, K. A., Frey, M. C., and Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence*, 36(2):153–160.

Kuncel, N. R., Crede, M., and 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.

Liao, H.-Y., Armstrong, P. I., and Rounds, J. (2008). Development and initial validation of public domain Basic Interest Markers. *Journal of Vocational Behavior*, 73(1):159–183.

Lord, F. M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, 20(1):1–22.

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

McCaffrey, R. J. and Lynch, J. K. (2009). Test security in the new millennium: Is this really psychology's problem? *Emerging Fields*, 21(2):27.

Murphy, L. L., Geisinger, K. F., Carlson, J. F., and Spies, R. A. (2011). *Tests in Print VIII.* An Index to Tests, Test Reviews, and the Literature on Specific Tests. Buros Institute of Mental Measurements, Lincoln, Nebraska.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 200:1–66.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Revelle, W. (2013). *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University, Evanston, Illinois. R package version 1.3.9.13.

Revelle, W. and Brown, A. (2013). Standard errors for SAPA correlations. In *Society for Multivariate Experimental Psychology*, pages 1–1, St. Petersburg, FL.

Revelle, W. and Condon, D. M. (2012). *Estimating ability for two samples.* Retrieved November 1, 2013, from http://www.personality-project.org/revelle/publications/EstimatingAbility.pdf.

Revelle, W., Wilt, J., and Rosenthal, A. (2010). Individual differences in cognition: New methods for examining the personality-cognition link. In Gruszka, A., Matthews, G., and Szymura, B., editors, *Handbook of Individual Differences in Cognition: Attention, Memory and Executive Control*, chapter 2, pages 27–49. Springer, New York.

Revelle, W. and Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74(1):145–154.

Sackett, P. R. and Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, 85(1):112–118.

Shipley, W. C., Gruber, C., Martin, T., and Klein, A. M. (2010). *Shipley Institute of Living Scale, 2nd edition*. Western Psychological Services, Los Angeles, CA.

Shipley, W. C., Gruber, C. P., Martin, T. A., and Klein, A. M. (2009). *Shipley-2*. Western Psychological Services, Los Angeles, CA.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. John Wiley & Sons Inc, London.

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

Tucker, L. R. and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1):1–10.

Uebersax, J. S. (2000). *Estimating a latent trait model by factor analysis of tetrachoric correlations.* Retrieved September 13, 2013, from http://www.john-uebersax.com/stat/irt.htm.

Wilt, J., Condon, D. M., and Revelle, W. (2011). Telemetrics and online data collection: Collecting data at a distance. In Laursen, B., Little, T. D., and Card, N. A., editors, *Handbook of Developmental Research Methods*, chapter 10, pages 163–180. Guilford Press.

Zinbarg, R. E., Revelle, W., Yovel, I., and Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1):123–133.

Zinbarg, R. E., Yovel, I., Revelle, W., and McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for omega hierarchical. *Applied Psychological Measurement*, 30(2):121–144.

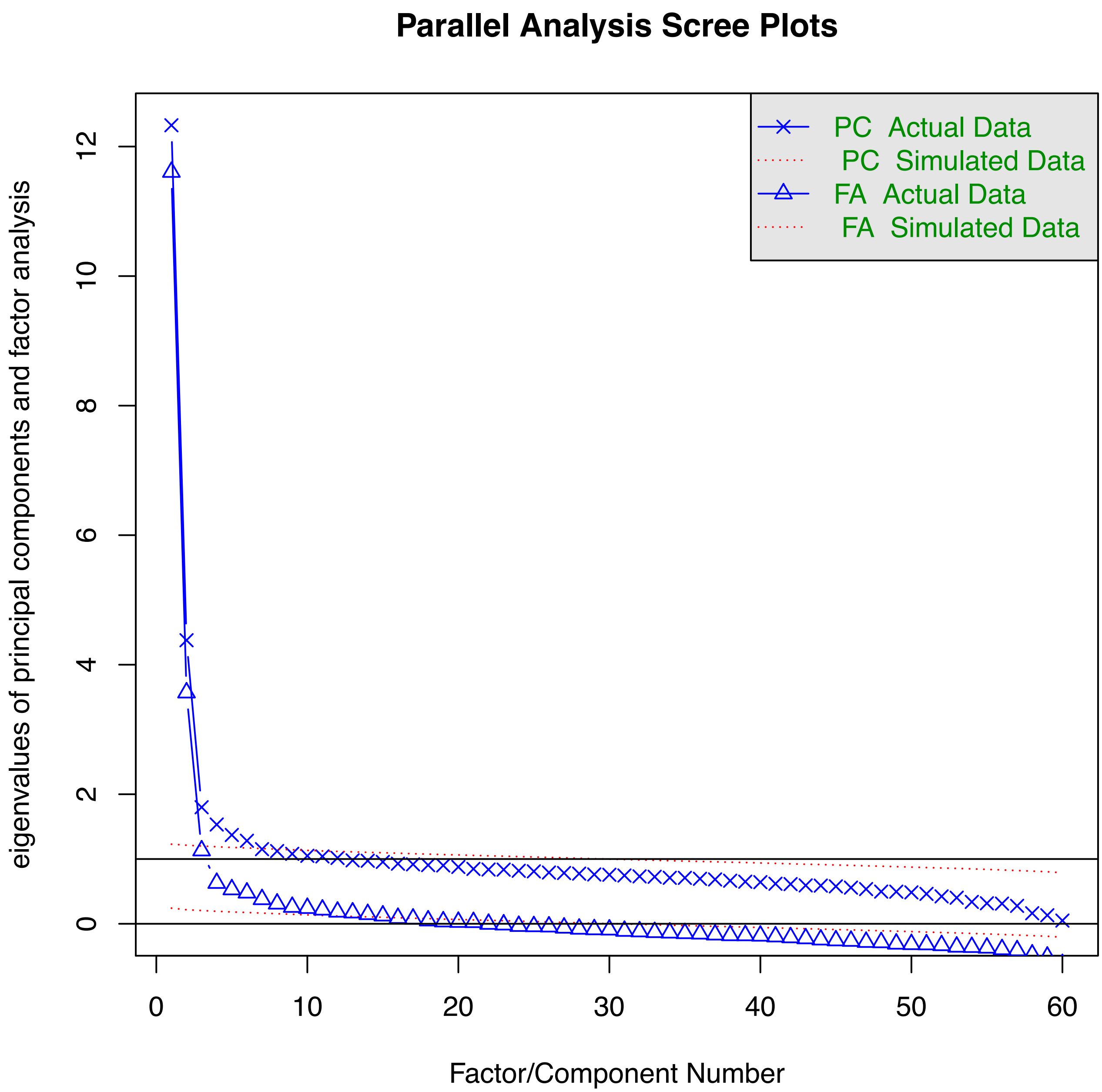


Figure 1: Scree plots based on all 60 ICAR items

Figure 2: Omega hierarchical for the *ICAR Sample Test*

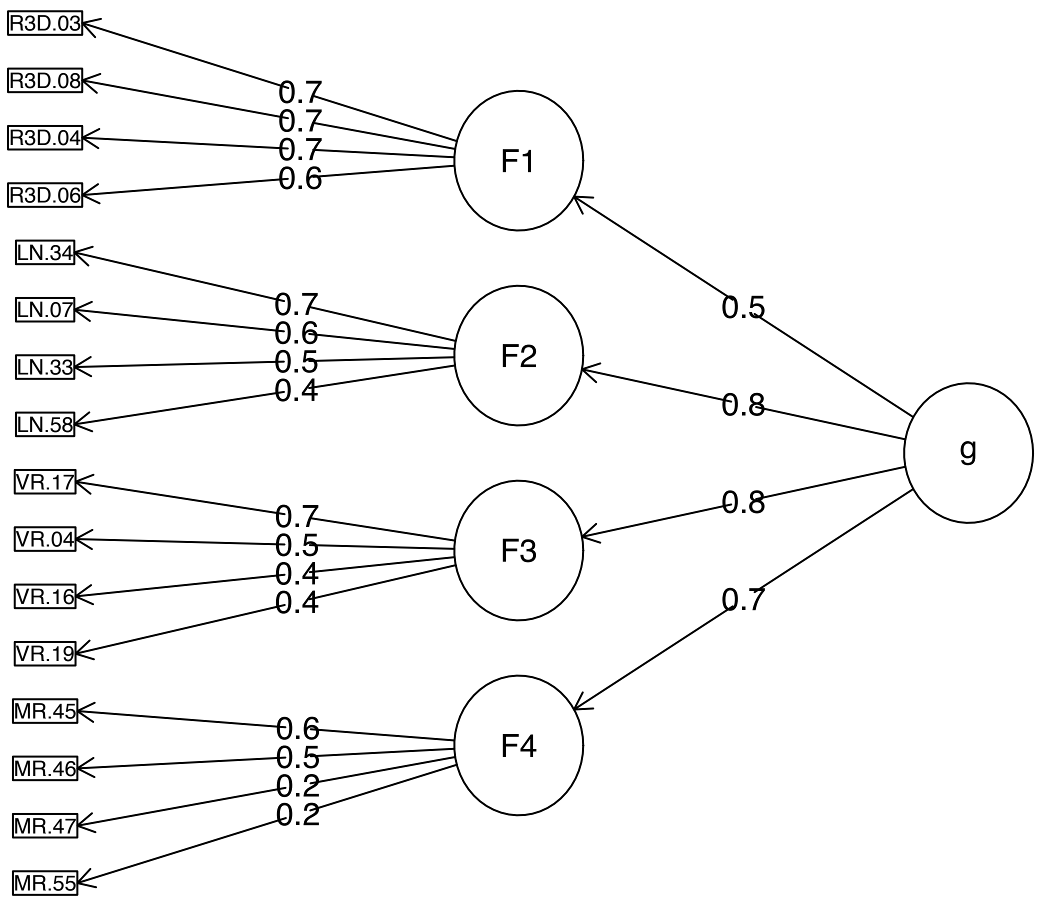
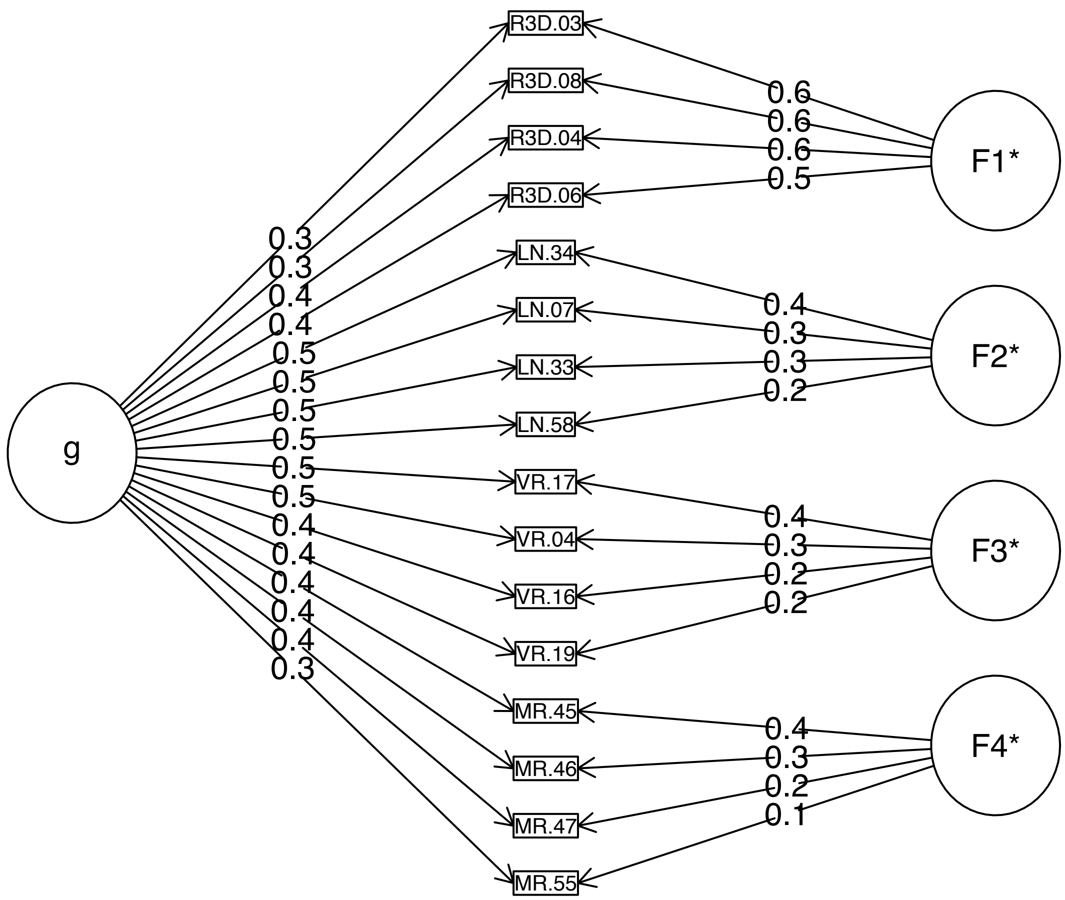


Figure 3: Omega with Schmid-Leiman transformation for the *ICAR Sample Test*



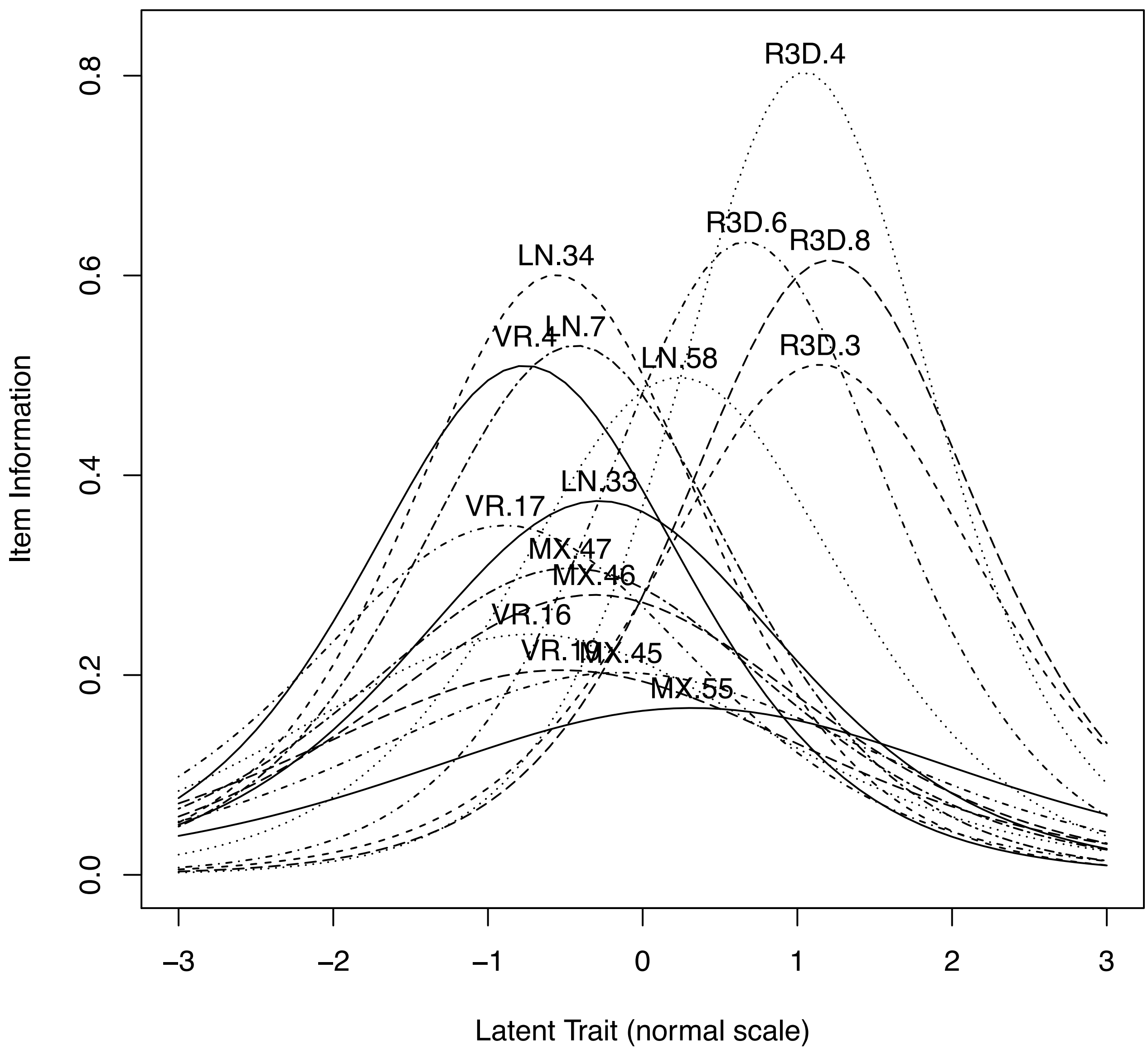


Figure 4: Item Information Functions for the 16 item *ICAR Sample Test*

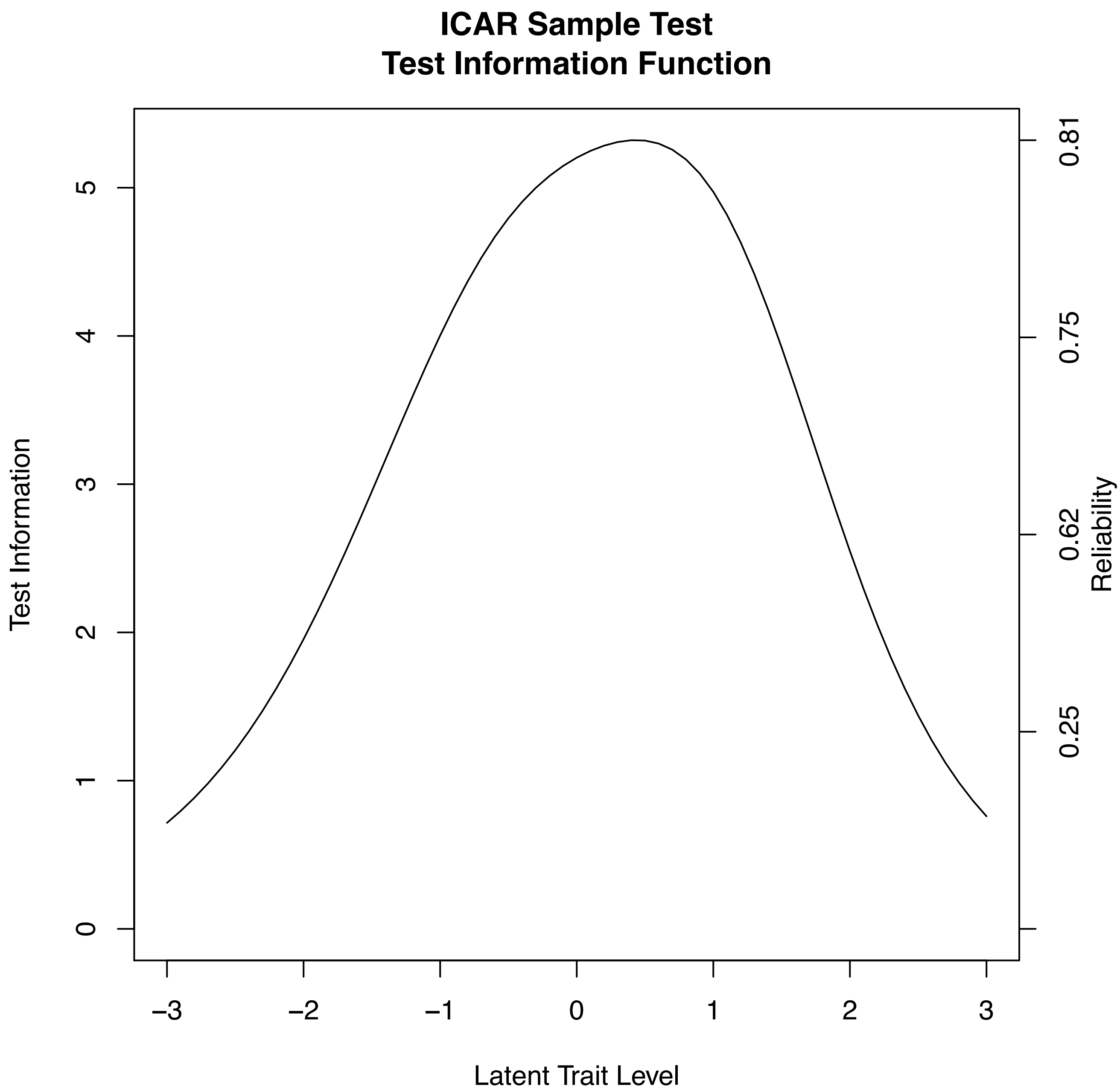


Figure 5: Test Information Function for the 16 item *ICAR Sample Test*

Table 1: Study 1 Participants by Educational Attainment

Educational attainment	% of total	Mean age	Median age
Less than 12 years	14.5%	17.3	17
High school graduate	6.2%	23.7	18
Currently in college/university	51.4%	24.2	21
Some college/university, but did not graduate	5.0%	33.2	30
College/university degree	11.7%	33.2	30
Currently in graduate or professional school	4.4%	30.0	27
Graduate or professional school degree	6.9%	38.6	36

Table 2: Descriptive statistics for the ICAR items administered in Study 1

Item	*n*	*mean*	*sd*	Item	*n*	*mean*	*sd*
LN.01	31,239	0.79	0.41	R3D.11	7,165	0.09	0.29
LN.03	31,173	0.59	0.49	R3D.12	7,168	0.13	0.34
LN.05	31,486	0.75	0.43	R3D.13	7,291	0.10	0.30
LN.06	34,097	0.46	0.50	R3D.14	7,185	0.14	0.35
LN.07	*36,346*	*0.62*	*0.49*	R3D.15	7,115	0.22	0.42
LN.33	*39,384*	*0.59*	*0.49*	R3D.16	7,241	0.30	0.46
LN.34	*36,655*	*0.62*	*0.48*	R3D.17	7,085	0.15	0.36
LN.35	34,372	0.47	0.50	R3D.18	6,988	0.13	0.34
LN.58	*39,047*	*0.42*	*0.49*	R3D.19	7,103	0.16	0.37
MR.43	29,812	0.77	0.42	R3D.20	7,203	0.39	0.49
MR.44	17,389	0.66	0.47	R3D.21	7,133	0.08	0.28
MR.45	*24,689*	*0.52*	*0.50*	R3D.22	7,369	0.30	0.46
MR.46	*34,952*	*0.60*	*0.49*	R3D.23	7,210	0.19	0.39
MR.47	*34,467*	*0.62*	*0.48*	R3D.24	7,000	0.19	0.39
MR.48	17,450	0.53	0.50	*VR.04*	*29,975*	*0.67*	*0.47*
MR.50	19,155	0.28	0.45	VR.09	25,402	0.70	0.46
MR.53	29,548	0.61	0.49	VR.11	26,644	0.86	0.35
MR.54	19,246	0.39	0.49	VR.13	24,147	0.24	0.43
MR.55	*24,430*	*0.36*	*0.48*	VR.14	26,100	0.74	0.44
MR.56	19,380	0.40	0.49	*VR.16*	*31,727*	*0.69*	*0.46*
R3D.01	7,537	0.08	0.28	*VR.17*	*31,552*	*0.73*	*0.44*
R3D.02	7,473	0.16	0.37	VR.18	26,474	0.96	0.20
R3D.03	*12,701*	*0.17*	*0.37*	*VR.19*	*30,556*	*0.61*	*0.49*
R3D.04	*12,959*	*0.21*	*0.41*	VR.23	24,928	0.27	0.44
R3D.05	7,526	0.24	0.43	VR.26	13,108	0.38	0.49
R3D.06	*12,894*	*0.29*	*0.46*	VR.31	26,272	0.90	0.30
R3D.07	7,745	0.12	0.33	VR.32	25,419	0.55	0.50
R3D.08	*12,973*	*0.17*	*0.37*	VR.36	25,076	0.40	0.49
R3D.09	7,244	0.28	0.45	VR.39	26,433	0.91	0.28
R3D.10	7,350	0.14	0.35	VR.42	25,108	0.66	0.47

Note: "LN" denotes Letter and Number Series, "MR" is Matrix Reasoning, "R3D" is Three-Dimensional Rotation, and "VR" is Verbal Reasoning. Italicized items denote those included in the 16-Item *ICAR Sample Test*.

Table 3: Alpha and omega for the ICAR item types

	α	ω_h	ω_t	items
ICAR60	0.93	0.61	0.94	60
LN items	0.77	0.66	0.80	9
MR items	0.68	0.58	0.71	11
R3D items	0.93	0.78	0.94	24
VR items	0.76	0.64	0.77	16
ICAR16	0.81	0.66	0.83	16

Note: ω_h = omega hierarchical, ω_t = omega total. Values are based on composites of Pearson correlations between items.

Table 4: Four-factor item loadings for the *ICAR Sample Test*

Item	Factor 1	Factor 2	Factor 3	Factor 4
R3D.03	**0.69**	-0.02	-0.04	0.01
R3D.08	**0.67**	-0.04	-0.01	0.02
R3D.04	**0.66**	0.03	0.01	0.00
R3D.06	**0.59**	0.06	0.07	-0.02
LN.34	-0.01	**0.68**	-0.01	-0.02
LN.07	-0.03	**0.60**	-0.01	0.05
LN.33	0.04	**0.52**	0.01	0.00
LN.58	0.08	**0.43**	0.07	0.01
VR.17	-0.04	0.00	**0.65**	-0.02
VR.04	0.06	-0.01	**0.51**	0.05
VR.16	0.02	0.05	**0.41**	0.00
VR.19	0.03	0.02	**0.38**	0.06
MR.45	-0.02	-0.01	0.01	**0.56**
MR.46	0.02	0.02	0.01	**0.50**
MR.47	0.05	0.18	0.10	**0.24**
MR.55	0.14	0.09	-0.04	**0.21**

Table 5: Correlations between factors for the *ICAR Sample Test*

	R3D Factor	LN Factor	VR Factor	MR Factor
R3D Factor	1.00			
LN Factor	0.44	1.00		
VR Factor	0.70	0.45	1.00	
MR Factor	0.63	0.41	0.59	1.00

Note: R3D = Three-Dimensional Rotation, LN = Letter and Number Series, VR = Verbal Reasoning, MR = Matrix Reasoning

Table 6: Item and test information for the 16 item *ICAR Sample Test*

	Latent Trait Level (normal scale)						
Item	-3	-2	-1	0	1	2	3
VR.04	0.07	0.23	0.49	0.42	0.16	0.04	0.01
VR.16	0.08	0.17	0.25	0.23	0.13	0.06	0.02
VR.17	0.09	0.27	0.46	0.34	0.13	0.04	0.01
VR.19	0.07	0.14	0.24	0.25	0.16	0.07	0.03
LN.07	0.06	0.18	0.38	0.39	0.19	0.06	0.02
LN.33	0.05	0.15	0.32	0.37	0.21	0.08	0.02
LN.34	0.05	0.20	0.46	0.45	0.19	0.05	0.01
LN.58	0.03	0.09	0.26	0.43	0.32	0.13	0.04
MR.45	0.05	0.11	0.17	0.20	0.16	0.09	0.04
MR.46	0.06	0.13	0.22	0.24	0.17	0.08	0.04
MR.47	0.06	0.16	0.31	0.32	0.18	0.07	0.02
MR.55	0.04	0.07	0.11	0.14	0.13	0.10	0.06
R3D.03	0.00	0.01	0.06	0.27	0.64	0.47	0.14
R3D.04	0.00	0.01	0.07	0.35	0.83	0.45	0.10
R3D.06	0.00	0.03	0.14	0.53	0.73	0.26	0.05
R3D.08	0.00	0.01	0.06	0.26	0.64	0.48	0.14
TIF	0.72	1.95	4.00	5.20	4.97	2.55	0.76
SEM	1.18	0.72	0.50	0.44	0.45	0.63	1.15
Reliability	NA	0.49	0.75	0.81	0.80	0.61	NA

Table 7: Self-reported achievement test scores and national norms

	Study 2 self-reported			published norms	
	n	mean	s.d.	mean	s.d.
SAT - Critical Reading	7,404	609	120	496	114
SAT - Math	7,453	611	121	514	117
ACT	12,254	25.4	5.0	21.1	5.2

Note: SAT norms are from the 2012 *Total Group Profile Report*. ACT norms are from the 2011 *ACT Profile Report*.

Table 8: Correlations between self-reported achievement test scores and ICAR composite scales

						ICAR composite scale scores				
	SAT-CR	SAT-M	SAT-CR+M	ACT	ICAR60	LN	MR	R3D	VR	ICAR16
SAT-CR[1]	*0.86*	0.83		0.69	0.52	0.41	0.37	0.39	0.68	0.52
SAT-M[2]	0.72	*0.88*		0.66	0.60	0.50	0.47	0.49	0.67	0.59
SAT-CR+M[3]			*0.89*	0.71	0.59	0.48	0.44	0.47	0.72	0.59
ACT[4]	0.62	0.60	0.65	*0.95*	0.52	0.39	0.35	0.44	0.61	0.52
ICAR60[5]	0.46	0.54	0.54	0.49	*0.93*					
LN[5]	0.33	0.41	0.40	0.33		*0.77*	0.84	0.59	0.90	
MR[5]	0.28	0.36	0.34	0.28		0.61	*0.68*	0.67	0.81	
R3D[5]	0.35	0.44	0.43	0.41		0.50	0.53	*0.93*	0.58	
VR[5]	0.55	0.55	0.59	0.52		0.69	0.58	0.49	*0.76*	
ICAR16[5]	0.43	0.50	0.50	0.46						*0.81*

Note: Uncorrected correlations below the diagonal, correlations corrected for reliability above the diagonal. Reliability values shown on the diagonal.

[1] n = 7,404

[2] n = 7,453

[3] n = 7,348

[4] n = 12,254

[5] Composite scales formed based on item correlations across the full sample (n = 34,229).

Table 9: Correlations between self-reported achievement test scores and IRT-based ICAR scores

					ICAR IRT-based scores				
	SAT-CR	SAT-M	SAT-CR+M	ACT	ICAR60	LN	MR	R3D	VR
SAT-CR[1]					0.44	0.37	0.35	0.37	0.44
SAT-M[2]	0.72				0.44	0.33	0.29	0.35	0.39
SAT-CR+M[3]	0.93	0.93			0.47	0.37	0.33	0.38	0.45
ACT[4]	0.62	0.60	0.65		0.44	0.35	0.32	0.38	0.43
ICAR60[5]	0.36	0.42	0.42	0.39					
LN[5]	0.24	0.28	0.28	0.24					
MR[5]	0.18	0.22	0.21	0.18		0.30			
R3D[5]	0.25	0.32	0.30	0.28		0.26	0.23		
VR[5]	0.35	0.36	0.38	0.36		0.36	0.26	0.22	

Note: IRT scores for ICAR measures based on 2 to 4 responses per participant for each item type (LN, MR, R3D, VR) and 12 to 16 responses for ICAR60. Uncorrected correlations are below the diagonal, correlations corrected for incidental selection are above the diagonal.

[1] $n = 7{,}404$

[2] $n = 7{,}453$

[3] $n = 7{,}348$

[4] $n = 12{,}254$

[5] $n = 34{,}229$

Table 10: Correlations between mean SAT norms, mean SAT scores in Study 2 and mean IRT-based ICAR scores when ranked by university major

	College Board Norms			Study 2 Self-Reported			Study 2 IRT-based			
	SAT-CR	SAT-M	SAT-CR+M	SAT-CR	SAT-M	SAT-CR+M	ICAR60	LN	MR	R3D
SAT-M norms	0.66									
SAT-CR+M norms	0.91	0.91								
SAT-CR study 2	0.79	0.61	0.77							
SAT-M study 2	0.56	0.80	0.74	0.81						
SAT-CR+M study 2	0.71	0.74	0.80	0.95	0.95					
ICAR60 study 2	0.53	0.84	0.75	0.60	0.77	0.72				
LN study 2	0.41	0.80	0.66	0.49	0.76	0.66	0.96			
MR study 2	0.22	0.66	0.48	0.23	0.52	0.39	0.83	0.78		
R3D study 2	0.42	0.80	0.67	0.50	0.71	0.64	0.94	0.92	0.82	
VR study 2	0.69	0.79	0.81	0.76	0.80	0.82	0.91	0.82	0.64	0.76

Note: $n = 29$.

Table 11: Correlations between mean GRE norms and mean IRT-based ICAR scores when ranked by university major

	ETS Norms			Study 2 IRT-based			
	GREV	GREQ	GREVQ	ICAR60	LN	MR	R3D
GREQ norms	0.23						
GREVQ norms	0.63	0.90					
ICAR60 study 2	0.54	0.78	0.86				
LN study 2	0.41	0.72	0.76	0.93			
MR study 2	0.42	0.71	0.75	0.86	0.81		
R3D study 2	0.44	0.80	0.83	0.92	0.86	0.75	
VR study 2	0.67	0.63	0.80	0.92	0.80	0.79	0.77

Note: $n = 91$.

Table 12: Standard deviations of scores for the unrestricted samples and Study 3

Sample	Block Patterns	Abstraction	*Shipley-2* Vocabulary	Composite A	Composite B	*ICAR Sample Test*
Unrestricted	15.0	15.0	15.0	15.0	15.0	1.86
Study 3	11.1	9.8	6.8	6.8	8.9	1.48

Note: Unrestricted standard deviations based on the published norms for the *Shipley-2* and the Study 1 sample for the *ICAR Sample Test*.

Table 13: Correlations between the *ICAR Sample Test* and the *Shipley-2*

ICAR16	Block Patterns[1]	Abstraction[2]	Vocabulary[3]	Composite A[2]	Composite B[1]
Uncorrected	0.40	0.44	0.15	0.41	0.41
Range corrected	0.64	0.69	0.59	0.68	0.68
Range & reliability corrected				0.82	0.81

[1] $n = 68$
[2] $n = 69$
[3] $n = 137$