

**Evaluating the Concurrent Validity of Three Web-Based IQ Tests and the
Reynolds Intellectual Assessment Scales (RIAS)**

Michael Firmin
Cedarville University

Chi-en Hwang
Cedarville University

Amanda Burger
Cedarville University

Jessica Sammons
University of Cincinnati

Ruth Lowrie
Clark County Educational Service Center

Abstract

In a double-blind study, 60 General Psychology students, selected in low, average, and high ACT ranges, were administered the Reynolds Intellectual Assessment Scales (RIAS). On a separate occasion, the students also completed web-based internet IQ tests from tickle.com, queendom.com, and iqtest.com. MANOVA results showed that ACT level had a significant effect on all four IQ scores combined (Wilk's Lambda = .451, $p < .001$), and all univariate ANOVA results also found significant ACT effects ($p < .001$ for RIAS, Tickle and Queendom; $p < .05$ for IQtest). Concurrent validity was evaluated by correlating scores on a particular test with the Composite Intelligence Index (CIX) scores on RIAS. Both Tickle and Queendom scores correlated significantly ($p < .01$) with RIAS, but the correlation between IQtest and RIAS was not significant. Correlations among the three internet tests were all significant ($p < .01$).

Evaluating the Concurrent Validity of Three Web-Based IQ Tests and the Reynolds Intellectual Assessment Scales (RIAS)

Since modern IQ testing began with Albert Binet and Theodore Simone in 1905 with the development of the Binet Intelligence Test, assessing intellectual ability has become an application of science, theory, and also a clinical business. Testing for mental retardation, gifted programs, school placement, learning disabilities, vocational assessment, neuropsychological examinations, and military placement are only a few routine milieus in which various forms of intellectual assessment occur. Boake (2002) suggests that since the days of WWI Army Alpha and Beta appraisals, the protocol of intelligence testing has undergone some refinements, but essentially little change has actually occurred.

Esters and Ittenbach (1997) argue that present-day IQ tests are not purely empirical instruments. Rather, constructs that the author makes regarding the nature of intelligence tend to determine the content of the assessment. The construct of “g” or global intellectual capacity observed and measurable across multiple life domains, for example, is a longstanding theory behind many present-day IQ tests (Sternberg, 2000). Although some have tried to break from the longstanding traditions of IQ testing with ideas such as Emotional Intelligence (Goleman, 1995), most psychologists are relatively rank-and-file with how testing should operate, even if variability exists among the theoretical constructs of intelligence.

One reason for the decades of stability that intellectual testing has shown relates to the standards imposed by psychologists. When Charles Spearman received his Ph.D. under the first experimental psychologist, Wilhelm Wundt, he set in motion a concept that provided the most salient anchor for tests of his generation and the rest to follow (DuBois, 1970). That concept was reliability. In order for an instrument to receive scientific sanction, and consequently psychological approval, it must be shown to produce similar results with each administration. It is no understatement to say that this is the rockbed principle of all psychological testing, including intellectual (Gregory, 1996).

The second psychometric anchoring principle is validity (McIntire & Miller, 2000). Essentially, a test must aptly measure what it purports to measure. In the case of IQ tests, an instrument must show measures of construct, concurrent, content, criterion, and/or convergent validity in order to find acceptance in the psychological world and to be given serious credence (Thorndike, 1997).

A third psychometric principle that anchors the testing process in psychology is standardization (Murphy & Davidshofer, 2001). This benchmark requires that tests be administered under the same conditions under which the norming process occurred. Thus, the results of tests administered by psychologists in various settings may be considered equivalent.

Psychologists are so committed to these principles of reliability, validity, and standardization that the concepts are embedded into the ethics codes of the American Psychological Association (2002). Using tests, including intellectual assessments, that do not meet acceptable thresholds of reliability and validity under standardized conditions for professional purposes such as diagnosing psychiatric or learning disorders is considered

unethical. When a school or clinical psychologist's license is placed at stake, the odds are high that psychologists will conform to the standards, which strengthens the stability of testing protocol, making change less likely to occur.

Having established the longstanding character of intelligence, its resistance to change, and factors that provide this stability—we suggest that a present potential threat exists to this solidarity. In the early developmental stages of what has become the modern computer revolution, it became clear that technology would make its attempt to capture part of the market's corner of intellectual and other means of assessment (Garnett, 1985). In academe, attempts were made to implement technology into the assessment process, holding to the principles of reliability, validity, and standardization (Ellis, 1991).

But then, outside of the control of academics and professional psychologists, the World Wide Web made its impact on home computers and accessibility to non-university life. This technology brought an economy of business and advertisement to anybody with a port hook-up and internet server. Moreover, virtually anybody with computer savvy could establish a web page and consequently sell goods. Thus, the market now became open to a wide variety of persons, some professional holding to the principles of reliability, validity, and standardization—and some not. Web merchants have found means to sell countless products to customers willing to pay for the services. This includes IQ tests.

No research was located vis-à-vis studies in the literature relating IQ tests offered on the internet to the three fundamental principles of psychological testing (reliability, validity, and standardization). Consequently, this domain of research appears to be virgin with multiple needs for empirical inquiry. The present research is not designed to answer the ultimate questions in this area, but rather to make an indent regarding to what degree such tests are psychometrically valid. An instrument can be reliable, but not valid—however, if it is invalid—then it is not reliable.

Philosophical questions relating to the use of web-based IQ tests are left for another article. The present study used one of many solid measurement instruments of intellectual ability as a comparison standard for three (3) internet IQ tests. Concurrent validity is established by showing high correlations between a test in question with currently used instruments which have shown solid psychometric properties in development and use. In sum, in the present study, we selected the Reynolds Intellectual Assessment Scales (RIAS) as a potential instrument for assessing the concurrent validity of Tickle, Queendom, and IQ Test, three popular and well advertised internet tests purporting to measure intellectual ability.

Method

Participants and Procedure

This study utilized 60 general psychology students from a private, comprehensive, Midwestern university. The students were randomly selected from three general psychology classes. Since the general psychology course is part of the general education curriculum at this

institution, the students represented a relative cross-section of the student population, with multiple majors represented.

All the students in the classes were placed into one of three categories according to their composite ACT scores, obtained from university records. Student composite SAT scores were converted to ACT equivalents, when needed, using a conversion chart obtained from the university admissions department and which is considered relatively standard for university admissions departments. The university accepts both ACT and SAT scores for admission, although the ACT is preferred. Also, if students had taken both the ACT and SAT tests, then we used the higher of the two reported scores as we considered the most accurate reflection of students' true aptitude. That is, one is more likely to show a false low score than a false high score.

The scores from participants in our sample ranged from 17 to 32. Given the student scores in our sample, and that the average ACT at the institution was 26, we utilized the following classifications for three groups: Low (24 and under), medium (25-27), and high (28 and above), with equal distributions of students in each category. The students were not informed of their level and never knew that this was a variable being considered in the study. In addition, the persons conducting the IQ tests did not know student ACT scores when administering the tests. Consequently, from that perspective the study was conducted double-blind.

Twenty 20 students were randomly selected from each of the three categories. That is, 20 students randomly were selected from the low category, 20 from the medium category, and 20 from the high category for a total of 60 student participants in the study. The rationale for utilizing the three groups was to ensure a reasonable distribution of academic ability among the subjects being studied. In other words, if all the subjects were from only one or two ability groups (e.g., high or low), then this could negatively affect the study's external validity. I is believed that by obtaining an equal sample across a range of ability levels, potential skew was avoided at the outset of the experimental design.

It is acknowledged that ACT/SAT scores are not measures of intellectual ability. It is not the intent to represent them in that manner nor to enter the debate of how valid they are relative to bias issues. Rather, this was a clean method to ensure that at least a minimum-level of sampling distribution was garnered relative to a reasonable measure of aptitude. Since all students already had taken the ACT or SAT, and they did not know that they were assigned to a particular group (high, medium, or low), it provided the best classification means to us, given the objectives of our study.

The 60 students were administered the Reynolds Intellectual Assessment Scales (RIAS) under standard conditions. On a separate occasion, the students also completed web-based internet IQ tests from tickle.com, queendom.com, and iqtest.com. During the internet testing session, a proctor gave the same instructions to all the students and monitored the testing to ensure standardization.

Materials

Reynolds Intellectual Assessment Scales (RIAS). The RIAS is an individually administered intelligence test suitable for ages 3 years through 94 years. A two-subtest Verbal Intelligence Index (VIX) and a two-subtest Nonverbal Intelligence Index (NIX) are combined to form the Composite Intelligence Index (CIX). The RIAS was developed by Psychological Assessment Resources (PAR) to provide a psychometrically reliable and valid general intelligence score in a more efficient amount of time than similar, but more plenary, tests such as Stanford-Binet and Wechsler tests. Due to the test's relative brevity, requiring only 20 to 25 minutes to administer, the RIAS generally is appropriate for school, clinic, and private practice use. The RIAS has been used successfully for the evaluation of learning disabilities, mental retardation, giftedness, physical/orthopedic impairment, memory impairment, emotional disturbance, and research.

The internal consistency reliability based on Cronbach's alpha for the RIAS during standardization ranged from .94 to .98. The test-retest reliability ranged from .75 through .97 based upon various age groups. Correlations between RIAS and the WAIS III composite scores range from .61 through .79 with all but two of the correlations exceeding .70.

Queendom.com's classical intelligence test--2nd revision. The Classical Intelligence test is composed of 60 questions and generates one main score and six sub-scores. The internal consistency reliability (Cronbach's alpha) is reported at .88. Queendom's manual lists criterion validity for the relationship between IQ scores and level of formal education, academic performance rating, field of work, and position within a company.

Tickle.com's classic IQ test. The Classic IQ test is comprised of 40 questions utilizing four intelligence scales: mathematical, visual-spatial, linguistic, and logical. The IQ test was developed utilizing questions from Mensa Workout tests and the *Shipley Institute of Living Scale*. The internal consistency reliability (Cronbach's alpha) is reported at .81. No manual was available to the authors for the intelligence test at the time this article was written.

IQTest.com's intelligence test. The Intelligence Test contains 38 questions and utilizes 13 intelligence scales: arithmetic, algebraic, rote utilization, logical, visual apprehension, spatial skill, intuition, general knowledge, vocabulary, short term memory, spelling, geometric, and computational speed. No manual was available to the authors for the intelligence test at the time this article was written.

Results

Table 1 presents the means scores and standard deviations of the four measures of general intelligence. MANOVA results showed that ACT level had a significant effect on all four IQ scores combined (Wilk's Lambda = .451, $p < .001$), and all univariate ANOVA results also found significant ACT effects ($p < .001$ for RIAS, Tickle and Queendom; $p < .05$ for IQtest). These findings supported our initial decision of using ACT level to control sampling bias.

Table 1
Mean Scores and Standard Deviations for Measures of Intelligence by ACT Level

ACT Level	n	<u>Intelligence Measures</u>			
		RIAS	Tickle	Queendom	IQtest
Low (24 or lower)	20	105.15 (6.85)	120.90 (5.03)	110.60 (11.81)	135.85 (9.89)
Medium (25 to 27)	20	112.90 (6.63)	128.30 (4.80)	118.95 (10.99)	140.10 (9.74)
High (28 or up)	20	115.30 (8.16)	132.55 (4.77)	124.10 (7.25)	143.40 (8.65)
Total	60	111.12 (8.35)	127.25 (6.82)	117.88 (11.50)	139.78 (9.79)
Minimum		97	113	87	121
Maximum		130	140	136	165

Concurrent validity was evaluated by correlating scores on a particular test with the Composite Intelligence Index (CIX) scores on RIAS. Table 2 presents the Pearson's correlation coefficients. Both Tickle and Queendom scores correlated significantly ($p < .01$) with RIAS, but the correlation between IQtest and RIAS was not significant. Correlations among the three internet tests were all significant ($p < .01$).

Table 2
Correlations among Four Measures of General Intelligence

Measure	Tickle	Queendom	IQtest
RIAS	.51**	.43**	.20

** $p < .01$.

Discussion

The correlation values between RIAS and the three internet tests indicated low to modest concurrent validities of the internet tests. Validity of the IQtest is clearly questionable due to its low correlation with the RIAS. Although Tickle and Queendom seemed to measure a similar construct, they shared roughly 25% and 16% of shared variance with RIAS (respectively).

One would expect typical criterion-related, concurrent validity scores to be much higher relative to standard measures of comparison. For example, correlations between the composite RIAS and Wechsler Adult Intelligence Scale, Third Edition (WAIS-III) full scale IQ scores is .79 (Reynolds & Kamphaus, 2003). The correlation coefficient between the full scale WAIS-III and global composite Stanford-Binet, Fourth Edition (SB:FE) IQ scores is .88 (Wechsler, 2002). In short, the correlations of .51, .43, and .20 found in the present study are inadequate for soundly establishing concurrent validity with the RIAS instrument.

In addition to statistical comparisons, such as correlation coefficients, Thorndike (1997) also recommends visually inspecting scores for practical implications of differences between criterion tests and the test(s) under consideration. To that end, Table 2 identifies the mean differences between the composite RIAS scores and scores on the individual web-based tests. It is first noted that each of the web-based test averages were higher than the RIAS. That is, the web-based tests overestimated subjects' intelligence levels when compared to the RIAS.

Second, the RIAS index scores were scaled to have an average of 100 with a standard deviation of 15, the typical measurement for intelligence tests (Reynolds & Kamphaus, 2003). As seen in Table 2, IQtest exceeded or neared two standard deviation differences in mean scores with the RIAS. Tickle met or exceeded one full standard deviation difference in average scores with the RIAS. Queendom average scores were within one standard deviation of RIAS. In sum, an inspection of score differences suggest that practitioners would find it difficult to interchange or adequately compare scores obtained from the standardized, individually-administered RIAS IQ test with web-based IQ scores, particularly IQtest and Tickle.

Table 3
Mean Score Differences for Measures of Intelligence by ACT Level

ACT Level	n	<u>Average IQ Differences from RIAS</u>			
		RIAS	Tickle	Queendom	IQtest
Low (24 or lower)	20	105	+16	+5	+31
Medium (25 to 27)	20	113	+15	+6	+27
High (28 or up)	20	115	+17	+9	+28
Total	60	111	+16	+6	+28

Limitations and Future Research

The present study possesses several limitations affecting our conclusions to varying degrees. First, participants in our sample completed the RIAS in one sitting, followed by all three of the web-based tests in another sitting. Control was not implemented for potential order effects, as the web tests were administered in the same order on each administration. Consequently, practice effects or fatigue may have positively or negatively affected the results of the study.

Second, the subjects in the sample were college students from a private, Midwest institution. The sample consisted of Caucasians with little minority representation. As such, additional data is needed from larger cross-sections of the U.S. population, more closely matching census representations.

Third, this study assessed one measure of criterion-related validity: concurrent validity. Predictive validity is another important criterion-related validity measure that the present study did not address. In addition, content or construct validity was not assessed in the present study, although they also are important components for judging the ultimate soundness of web-based tests.

Finally, while the RIAS is a solid, well-recognized, and psychometrically validated instrument, it is not the gold standard SB-V or WAIS-III. In order to obtain more precise data regarding criterion-related, concurrent validity of web-based tests, additional research should be conducted. Comparing participant scores from more sophisticated tests will help to answer the questions posed in the present study with more accuracy.

References

- American Psychological Association. (2002). *Ethical principles of psychologists and standards of conduct*. Washington, D.C.: Author.
- Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology*, 24, 383-405.
- DuBois, P. H. (1970). *A history of psychological testing*. Boston: Allyn & Bacon.
- Ellis, C.R. (1991). The utility of a computerized assessment battery to evaluate cognitive functioning and attention. *Dissertation Abstracts International*, 52, (03), 1714. (UMI No. 9315947).
- Gregory, R. J. (1996). *Psychological testing: History, principles, and applications*. Boston: Allyn & Bacon.
- Esters, I.G., & Ittenbach, R.F. (1997). Today's IQ tests: Are they really better than their historical predecessors? *School Psychology Review*, 26, 211-224.
- Garnett, P.D. (1985). Intelligence measurement by computerized information processing using inspection time. *Journal of General Psychology*, 112, 325-335.
- Goleman, D. (1995). *Emotional intelligence*. New York: Bantam Books.
- McIntire, S. A. & Miller, L.A. (2000) *Foundations of psychological testing*. 2nd ed. Boston: Allyn & Bacon.
- Murphy, K. R. & Davidshofer, C. O. (2001). *Psychological testing*. Upper Saddle River, NJ: Prentice Hall.
- Reynolds, C. R. & Kamphaus. (2003). *Reynolds Intellectual Assessment Scales professional manual*. Lutz, FL: Psychological Assessment Resources.
- Sternberg, R.J. (2000). The holy grail of general intelligence. *Science*, 289, 399-401.
- Thorndike, R. M. (1997) *Measurement and evaluation in psychology and education*. 6th edition. Upper Saddle River, NJ: Prentice Hall.
- Wechsler, D. (2002). *WAIS-III and WMS-III technical manual*. New York: The Psychological Corporation.