IAP  AP 101

APPLIED PSYCHOMETRICS 101: IQ TEST SCORE DIFFERENCE SERIES

#1: Understanding global IQ test correlations

Despite reported evidence of strong concurrent correlations among IQ tests (concurrent validity), different IQ tests often produce different IQs for the same individual. This may be due to a number of factors. Prior to discussing the various factors, one must first understand the basic language of typical IQ-IQ comparison research. In the first of this series, IQ-IQ test correlations are explained. Statistically significant high correlations between different IQ tests, although providing strong concurrent validity evidence for tests, do not guarantee similar or identical IQs for all individuals tested.

Kevin S. McGrew, Ph.D.

Educational Psychologist

Director

Institute for Applied Psychometrics (IAP)

Often in clinical and forensic psychological testing, an individual will be administered different intelligence (IQ) tests by one or more examiners. Despite using tests with different content and standardization dates, the global (full scale) IQ from the different tests are frequently similar, or are reasonably close (when *measurement error* is taken into consideration). Other times the two IQs will be markedly different, a finding that often produces consternation for examiners and recipients of psychological reports.

Factors contributing to significant IQ differences are many, and include: (a) procedural or test administration issues (e.g., scoring errors; improper test administration; malingering; age vs grade norms), (b) test norm or standardization differences (e.g., possible errors in the norms; sampling plan for selecting subjects for developing the test norms; publication date of test), (c) content differences, and/or, (d) in the case of group research, research methodology issues (e.g., sample pre-selection effects on reported mean IQs) (McGrew, 1994).

Before addressing the potential reasons for IQ-IQ differences, users and consumers of IQ test results need to understand a basic psychometric concept used to quantify and describe the expected degree of correspondence between scores from different IQ tests. This brief report will provide an overview and explanation of one primary (and other related) psychometric concept—test score *correlations*.

The goal of the current **Applied Psychometrics 101** (**AP101**) report is to provide an understandable, conceptual, and statistically "lite" explanation of IQ-IQ correlations and expected differences.[1] Detailed mathematical and statistical nuances, advanced topics, and special circumstances are not discussed. Examples using data from a real sample are used to make the presentation as concrete as possible.

Subsequent issues in this **AP101 series** (*IQ Test Score Difference Series*) will attempt to address the major factors for differences (as listed above).

Correlation and related terms defined

The *APA Dictionary of Psychology* (VandenBos, 2007) defines *correlation*, *correlation coefficient*, and the *coefficient of determination* as:

> *Correlation:* "The degree of relationship (usually linear) between two attributes" (p. 234).
>
> *Correlation coefficient*: "A numerical index reflecting the degree of relationship (usually linear) between two attributes scaled so that the value of +1 indicates a perfect positive relationship, -1 a perfect negative relationship, and 0 no relationship" (p.234).
>
> *Coefficient of determination*. "A numerical index that reflects the degree to which variation in the dependent variable is accounted for by one independent variable. Also called **determination coefficient**" (p.186).

Correlation and correlation coefficient demonstrated

[1] The issue of an individual obtaining different IQ scores on the same test administered by different examiners is not the focus of the current report.

The correlations between different IQ tests are usually reported in each test's respective technical manual or in journal research reports. Typically two or more IQ tests are administered to all individuals in one or more research samples. In this **AP101** report, the 10th-11th grade concurrent validity sample of normal (above average) subjects described in the *WJ-R Technical Manual* (McGrew, Werder & Woodcock, 1991) is used. As described by McGrew et al. (1991), the sample consisted of 55 subjects who were administered both the *Wechsler Adult Intelligence Scale—Revised (WAIS-R*) and the *Woodcock-Johnson—Revised (WJ-R) Tests of Cognitive Ability*. As is usual with most major intelligence tests, both the WAIS-R and WJ-R provide IQs with an average score (*Mean*) of 100 and a standard deviation (*SD*) of 15.

Using samples that had been administered different sets of IQ tests, the same tests but newer revisions (WAIS-IV; WJ III), or samples of subjects of different ages, could just as well have been used to illustrate the same concepts. This research sample was selected for illustrative purposes since the mean IQs reported for the WAIS-R (118.5) and WJ-R (118.1; for BCA-Standard Scale) were nearly identical.

Visualization of sample data

Below is a *scatterplot* (Figure 1) of the WAIS-R and WJ-R global IQs for the 55 subjects. A scatterplot is a simple plot of the two respective scores (for each subject) on the X-axis (WAIS-R) and Y-axis (WJ-R) in a two dimensional graph. Each small oval or dot represents a person's IQ on the two respective IQ tests.

A quick inspection of the WAIS-R/WJ-R scatterplot below reveals a number of immediate conclusions.

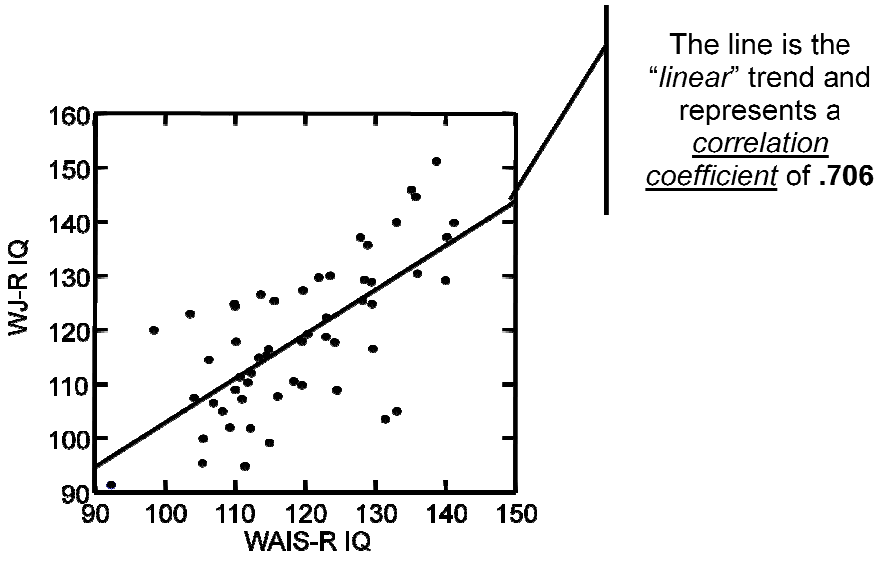


Figure 1: Scatterplot of WAIS-R and WJ-R IQs

- In general, as WAIS-R IQs increase so do the WJ-R IQs. This reflects a positive correlation.
- The correlation between the two IQs is statistically significant and high (r =.706) and is represented by the linear regression trend line in Figure 1.
- There is not a 1-1 correspondence between IQs for many subjects. Despite a high and statistically significant correlation (r =.706), which reflects the association between WAIS-R and

WJ-R IQs for the group sample, it is obvious that for a number of individual subjects, significant IQ-IQ discrepancies are present.

Interpretation of IQ-IQ correlations

Correlations reported between major IQ tests usually range from the .60s to .80s, with the highest correlations typically found in the .70 to .80 range (Kamphaus, 2005). Although statistically significant high correlations, it is important to recognize that these correlations estimate the relationship of two IQs across individuals (i.e., in the research sample group) and can lead to a false sense of expected IQ-IQ correspondence for individuals. This point is made concrete by converting the WAIS-R/WJ-R correlation of .706 to a more understandable metric.

Although correlations of .70 and .80 sound high, and are often touted as "high" in test manuals, research reports, and by experts,[2] correlations of this magnitude may inadvertently hide the "real world' reality of the similarities and differences between the two IQ test scores for individuals. To better understand the pragmatic reality of IQ-IQ correlations we calculate the *coefficient of determination* (r^2). This index is obtained by squaring the obtained correlation (.706 x .706) and multiplying the result by 100. The result is 49.8. What is this statistic? This is *coefficient of determination* which quantifies the amount of *shared or common test score variance* between the two tests (Neisser et al., 1996; Sattler, 2001). In this example, the .706 correlation indicates that the WAIS-R and WJ-R IQs have 49.8 % *common or shared test score variance*.

What is the interpretation of the remaining 50% test variance that is not in common between the WAIS-R and WJ-R IQ test batteries? The interpretation is that the other half of the uncommon WAIS-R and WJ-R IQ variance is due to (a) different abilities being measured by the two different IQ tests, and (b) to a lesser extent, *measurement error* due to less than perfect *reliability* for each test score.

Visual inspection of Figure 1, when combined with the knowledge that the WAIS-R and WJ-R have approximately 50% shared and unshared IQ variance, should lead the reader to the conclusion that not all individuals will receive the same IQ (or nearly similar scores) on these two different IQ tests. If each IQ test measures 50% of the same abilities as the other IQ test, an individual may perform higher or lower (on either of the tests) due to their performance on the tests in the respective IQ batteries measuring *abilities not shared or measured in common*. Using a simple analogy, comparing the total WAIS-R and WJ-R scores is akin to comparing two scores that are: (a) both based on the addition of the same apples (50% common or shared variance) and, (b) then the addition of different other types of fruit to the respective two IQ tests (one test then adds in oranges while the other test adds in grapefruits).

Two other examples are provided to illustrate this point. In the WAIS-III technical manual (The Psychological Corporation, 1997) a correlation of .88 (statistically significant and high) is reported between the WAIS-III Full Scale IQ and the Stanford-Binet Intelligence Scale—Fourth Edition (SB4) global score (n = 26 adult subjects). A correlation of .65 is also reported between the WAIS-III IQ and the special purpose Raven's Standard Progressive Matrices (SPM) in the same sample. Correlations of this magnitude, when converted to *coefficients of determination*, indicate that the WAIS-III has 77% and 42% common or shared variance with the SB4 and SPM, respectively. The WAIS-III/SB4 77% value is high and impressive. Yet, again, it is important to recognize that this group study suggests that the

[2] These are high values for test validation purposes based on group sample research. "Touting" such high correlations from a group-based concurrent validity study in a journal or technical manual is appropriate. The concern stated here is the possible misunderstanding that may occur when such high correlations are used to support predictions about a specific individual, without providing associated information re: the amount of known error of prediction associated with the correlations.

WAIS-III and SB4 still have 23% (approximately 1/4) of their respective test score variance that they do not share in common. The WAIS-III and SPM have more they don't share (58%) than they do measure in common (42%).

The only time one can expect two different IQ tests to provide approximately the same IQs for all individuals is if the tests are nearly perfectly correlated (correlation approaches +1.0).[3] This is not the reality reflected by decades of psychometric IQ comparison research. Although typical group-based correlations reported between major IQ tests (.70s to .80s) may sound impressive to non-psychometricians, correlations of this magnitude suggest that different IQ tests measure approximately 50% to 60% common abilities.[4] Not to be misunderstood—these are very impressive values for group-based test validation purposes. With 40% to 50% of their respective score variance accounted for by the measurement of different abilities by the two IQ tests (and some reflecting measurement error for each IQ test) one cannot expect two different IQ tests to provide scores that are always similar for all tested individuals. Different IQs are to be expected with regularity.

How often, and by how much, should IQs from different IQ tests be expected?

IQ-IQ expected differences

Inspection of the previous WAIS-R/WJ-R scatterplot (Figure 1) reveals a number of subjects with large differences between their respective WAIS-R and WJ-R IQs. To investigate further, IQ differences were calculated by subtracting each subjects WJ-R IQ from the subjects respective WAIS-R IQ (i.e., WAIS-R IQ – WJ-R IQ = *IQ difference*). The largest positive and negative IQ differences are designated in the second scatterplot (Figure 2). It is obvious that even when two psychometrically sound IQ tests are highly correlated (e.g., .706) psychologists will be often face discrepant scores for individuals. Sometimes the IQ-IQ differences will be quite large.

In Figure 2, the largest positive (+27) and negative (-19) WAIS-R/WJ-R IQ differences are labeled. If we assume that the respective IQs are valid, discrepancies of this magnitude will require expert interpretation. More importantly, the presence of such discrepant IQs (and others included in Figure 2) indicates that in applied settings important decisions (e.g., eligibility for special services; eligibility for disability support and income; diagnosis of MR in death penalty cases) will not be easy if two different psychometrically sound IQ tests are administered to a single subject and highly discrepant IQs are obtained. This is particularly true if rules, guidelines, or decisions are based on strict absolute IQ point cut scores.[5]

Given any two different IQ tests, and assuming that research has reported correlations between the two tests, is it possible to determine the expected magnitude and frequency of IQ-IQ discrepancies? The answer is "yes."

---

[3] Technically two tests with a perfect correlation may not produce the same identical IQ scores for individuals a number of reasons. Near perfect to a perfect correlation of +1.0 is a necessary but not sufficient condition for this to occur. See additional information in "*Caution: Correlations reveal nothing about typical mean IQ score differences*" section of current document for addition information.

[4] There are more complex multivariate statistical methods (e.g., joint factor analysis; canonical correlation) that can provide better estimates of common variance and, more importantly, help understand what the two respective IQ tests measure in common. Discussing them at this time would only cloud the interpretation and understanding of this brief conceptual discussion.

[5] The use of absolute and strict single point eligibility or cut-scores is an issue that warrants a special report and will not be addressed in this document.

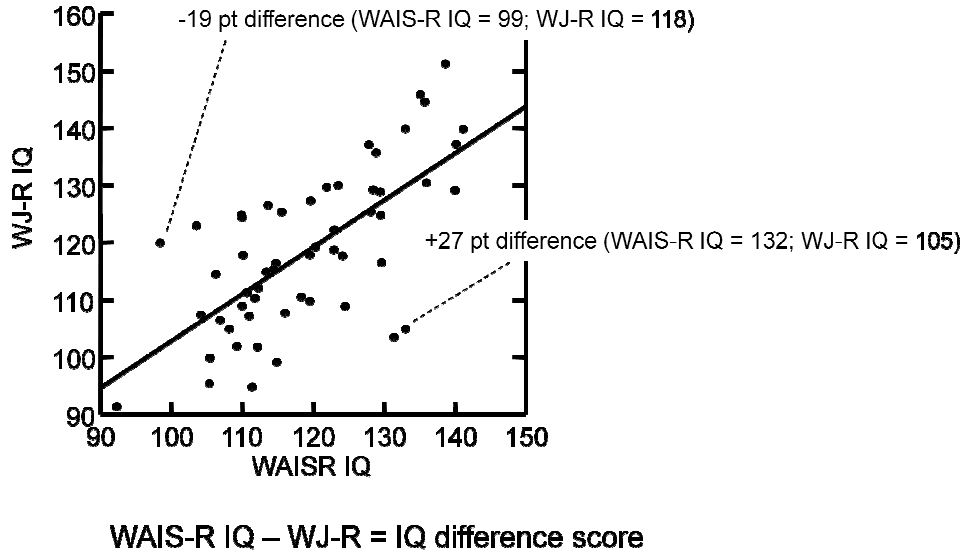


Figure 2: Scatterplot of WAIS-R and WJ-R IQs (Figure 1) with the two most extreme WAIS-R/WJ-R IQ Differences labeled.

For the WAIS-R/WJ-R sample of 55 subjects, the frequency of the IQ differences is displayed in a bar chart (Figure 3). It is obvious from inspection of Figure 3 that there is a considerable range of positive and negative WAIS-R/WJ-R IQ differences and, more importantly, IQ differences are roughly equally distributed around the average IQ difference of approximately zero (Mean IQ difference = 0.4). In terms of the spread of the sample IQ differences, approximately 68 % of the discrepancy scores range between -10 and +10 (sample IQ difference standard deviation = 9.9).

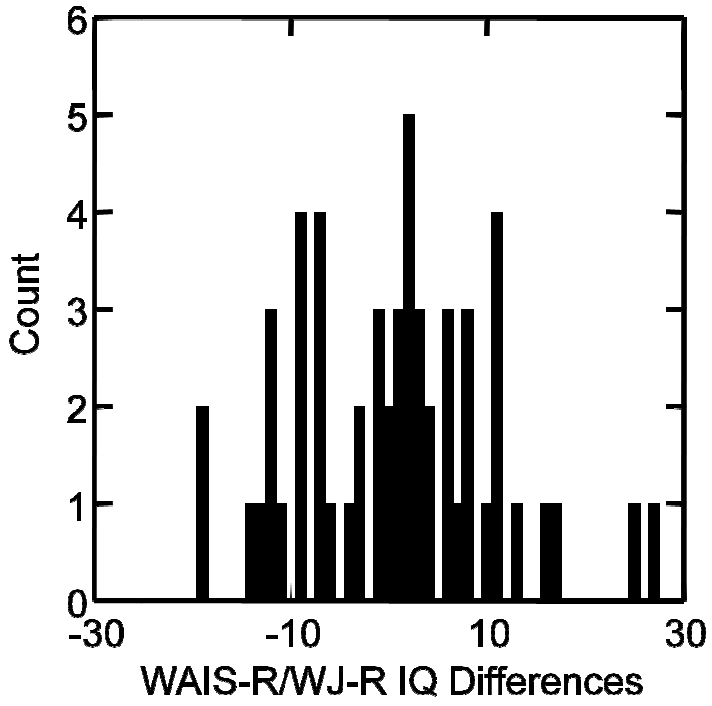


Figure 3: Bar chart of WAIS-R/WJ-R IQ differences

Are the above IQ difference findings unique to this sample of 55 subjects? No. More importantly, often the necessary statistics are available to estimate the distribution of IQ differences between different IQ tests via the use of a known psychometric equation.

Given any two correlated measures (IQs in the current discussion), if they are on a common scale (*Mean* =100; *SD* = 15; which is the case with most all major intelligence tests), a simplified formula can be used to calculate the expected *standard deviation (SD) of the discrepancies*.[6] Thus, all one needs is a solid estimate of the correlation (*r*) between the two IQ tests, a statistic that is often available from special validity studies reported in an IQ tests technical manual or in a synthesis of research studies published in professional journals. The formula is:

$$SD(diff) = 15 \times [\text{SQRT}(2 - 2 \times r)]$$

[Note: SQRT = square root]

Given the sample correlation of .706, if one did not have actual discrepancy scores to plot (as we did in the above example; see Figure 3), one would calculate the following:

[6] The complete formula for the *standard deviation of difference scores* is more complex than reported here but simplifies to the formula used in this report when the two correlated measures have the same standard deviation (*SD*). I would like to thank *Dr. Joel Schneider* for providing this simplified formula and pointing out potential confusion in the first draft of this report which discussed both the *SD*(diff) and the *standard error of estimate* [*SE(est)]* formula. When measures are highly correlated the two relevant formulas provide very similar results, although they are estimating different statistics (one the *SD* of the residuals of prediction and the other the *SD* of difference scores)

SE(diff) = 15 x [SQRT(2 - 2 x .706)]

SE(est) = **11.6**

The obtained value of 11.6 is reasonably close to the value calculated from analysis of the actual WAIS-R/WJ-R IQ differences in our sample data (9.9), thus demonstrating the validity of the formula.

Therefore, if professionals have good estimates of the correlations between different IQ tests, the above formula can be used to generate estimates of the *SD(diff)* which is the *standard deviation of the estimated IQ differences*. The value can be used to calculate the range of typical IQ-IQ difference scores (± 1 *SD*; 68 % of population should be between these values) that may be expected. These values are calculated for a number of typical IQ-IQ test correlations and are summarized in the Table 1.

IQ-IQ correlation	*SD* of estimated IQ-IQ differences [*SD(diff)*]	Range of typical (68%; ± 1 *SD*) IQ-IQ differences (based on whole numbers in prior column)
.60	13.4	26
.65	12.5	24
.70	11.6	22
.75	10.6	20
.80	9.5	18
.85	8.2	16

Table 1. *SD* of estimated IQ-IQ differences for different IQ-IQ correlations

Understanding the practical implications of the information in Table 1

The information presented in Table 1 can be used to understand the range of possible IQ-IQ differences for two different IQ tests when an estimate of the correlation between the two IQ tests is available. Using the information in the third column of Table 1, the conclusion is reached that **IQ-IQ differences will typically range (for 68% of the population) from as much as 26 (± 13) points (*r* = .60) to as little as 16 (± 8) points (*r* = .85).** Also, if published research is available documenting the average mean score difference between two different IQ tests, a slight change in calculating the "expected range of IQ-IQ differences" is demonstrated.

Example scenario: IQ-IQ tests correlate .70 and have average mean difference of 0.0

When two IQ tests are estimated to correlate .70 (based on data from technical manuals or professional journals), and if it is known (or assumed) that the two tests typically provide similar IQs (based on prior research), 68% (± 1 *SD*) of subjects administered both IQ tests will be expected to show IQ-IQ test differences that range from -11.6 to +11.6 points. Most will group around zero, but IQ differences of ± 5 points will not be rare, as well as IQ differences up to ± 11.6 points.

Example scenario: IQ-IQ tests correlate .70 and have average mean difference of 5 IQ points.

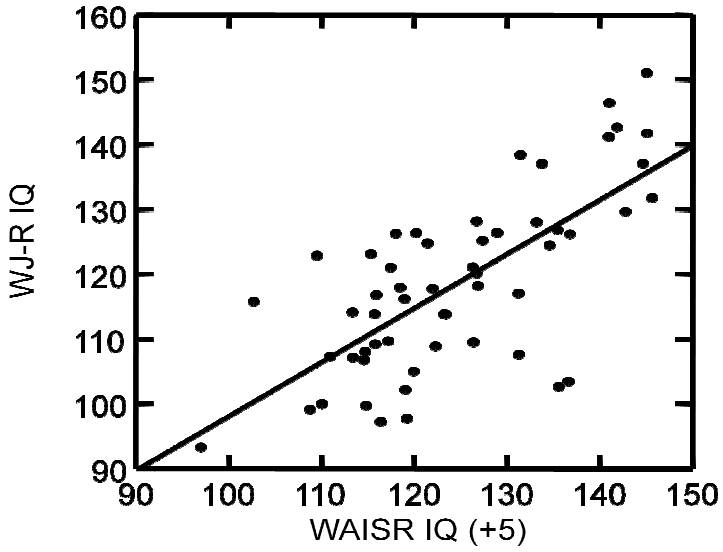
If two IQ tests correlate .70 (same as preceding example), but it is known (through a review of the literature) that IQ Test A (on the average) scores 5 points higher than IQ Test B (*average mean difference*)[7], when subjects are administered both IQ tests, the *expected mean difference* will be 5 points (average mean difference). IQ-IQ test differences will also range from -11.6 to +11.6 points. Given the 5 point expected difference (in favor of Test A scoring higher—when calculating the IQ-IQ difference by subtracting Test B from Test A) most IQ differences will group around +5 points, with 68% of the population expected to display a Test A-Test B IQ difference between -6.6 (5 – 11.6) and +16.6 (5 + 11.6) points.

Caution: Correlations reveal nothing about typical mean IQ differences

Correlations, no matter how high, provide no information regarding expected IQ mean differences between two IQ tests. Correlations only indicate the degree which two variables (e.g., two IQ tests) rank order subjects in roughly the same order. Two IQ tests could have a high correlation of .80. However, they may differ in major ways (content; data of publication and possible *Flynn Effect*; norm sample differences) that can produce systematically higher or low scores not reflected in correlations. This is best illustrated via a simulation with the sample data used above.

The WAIS-R IQs for all subjects in the sample were adjusted with a constant +5 IQ points. Five (5) IQ points was added to every subject's WAIS-R IQ. The new WAIS-R(+5) IQ was then correlated with the subjects WJ-R scores and a scatterplot generated (see Figure 4). When compared to the first scatterplot (Figures 1 and 2), the scatterplot in Figure 4 is identical with one exception—the linear trend line (which represents the correlation) has simply been shifted down by a constant of 5 points. The correlation is still .706.

The conclusion from this simulation proves what is known in the psychometric literature. Correlations only reveal the strength of the association of the relative ordering of subjects on two sets of scores. Correlations reveal no information about average (mean) scores. High correlations between IQ tests cannot be used to suggest that the tests should provide similar absolute scores for all individuals.



[7] In this example it is assumed that the *Flynn Effect* (Flynn, 2006; Neisser et al., 1996) is not operating. For this scenario it is assumed that both tests were published at approximately the same time and differ primarily in content coverage.

Figure 4: Scatterplot of WAIS-R (+5) and WJ-R IQs

Concluding comments

In an ever-changing world, psychological testing remains the flagship of applied psychology

(Embretson, 1996, p. 341).

There is little doubt that the psychometric measurement of human abilities, and intelligence testing in particular, stands as one of the major technical and scientific contributions that has emerged from the field of psychology. Reliable and valid measures of intelligence have demonstrated the greatest breadth of statistically significant moderate to high correlations with many other human traits and environmental outcomes (Neisser et al., 1996). As a result, "IQ tests" have become a cornerstone of the clinical and empirical study of human individual differences. They have become valuable practical tools in individual clinical assessment.

Unfortunately, IQ tests have also been surrounded by controversy (see Neisser et al., 1997; Sattler, 2001), much related to a misunderstanding of the strengths and limitations of the psychometric tools. Although among the best technical tools to emerge from the field of psychology, they are fallible (less than perfect). Many of the controversies and problems in the use of IQ tests have arisen from an exaggerated belief in their power, precision, and predictive capabilities, especially in the case of individual people and individual IQs. The purpose of this **Applied Psychometric 101** report was to provide an accurate appraisal of one characteristic of IQ tests in hopes of reducing the inappropriate use and interpretation of point-specific IQs when making critical decisions about individuals. In particular, this report has focused on understanding the "why" and "how often" of IQ-IQ differences.

Group-based research finds that most individually administered IQ tests correlate at statistically significant high levels (correlations ranging from .60 to .80). These group-based statistics are indeed impressive. However, even if two IQ tests are standardized at the same time, administered appropriately, and are psychometrically sound, although average (mean) group scores may be similar in research reports, psychologists (and recipients of psychological reports) must recognize that at the level of individuals, significant IQ-IQ differences will occur with regularity. The range of expected IQ-IQ differences for most of the population (68%; + 1 SD) will likely be 16 (+ or - 8) to 26 (+ or -13) IQ-IQ difference points. This reflects the current state-of-the-art of psychometric IQ testing. Current IQ technology does not allow for the assumption that all IQ tests will produce identical (or nearly identical) IQs for all individuals. Interpretation of psychometrically reliable and valid IQ-IQ test differences must be accepted and interpreted when they occur. Psychologists with appropriate expertise have a professional and ethical responsibility to seek out possible research-based explanations and hypotheses for why such differences may occur.

References

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

Flynn, J. (2006). TETHERING THE ELEPHAN: Capital Cases, IQ, and the Flynn Effect. *Psychology, Public Policy, and Law, 12(2)*, 170 –189

Kamphaus, R. (2005). *Clinical assessment of child and adolescent* intelligence (2nd Ed.). New York: Springer-Verlag.

McGrew, K. (1994). *Clinical interpretation of the Woodcock-Johnson Tests of Cognitive Ability-Revised.* Boston: Allyn and Bacon.

McGrew, K., Werder, J., & Woodcock, R. (1991). *WJ-R technical manual.* Chicago, IL: Riverside.

The Psychological Corporation, (1997). *WAIS-III, WMS-III technical manual*. San Antonio, TX: Author.

Neisser, U., Boodoo, G., Bouchard, T. J. Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. American Psychologist, 51(2), 77-101.

Sattler, J. (2001). *Assessment of children: Cognitive applications* (4th Ed.). La Mesa, CA: Jermoe M. Sattler, Publisher, Inc.

VandenBos, G. (2007). *APA Dictionary of Psychology*. Washington, DC: American Psychological Association.

Author information and conflict of interest disclosure

Dr. Kevin S. McGrew, Ph.D., is an Educational Psychologist with expertise and interests in applied psychometrics, intelligence theories and testing, human cognition, cognitive and non-cognitive individual difference variables impacting school learning, models of personal competence, conceptualization and measurement of adaptive behavior, measurement issues surrounding the assessment of individuals with disabilities, brain rhythm and mental timing research, and improving the use and understanding of psychological measurement and statistical information by professionals and the public. Prior to establishing IAP, Dr. McGrew was a practicing school psychologist for 12 years. McGrew received his Ph.D. in Educational Psychology (Special Education) from the University of Minnesota in 1989.

Dr. McGrew is currently Director of the *Institute for Applied Psychometrics* (IAP), a privately owned applied research organization established by McGrew. He is also the *Research Director for the Woodcock-Munoz Foundation* (WMF), Associate Director for *Measurement Learning Consultants* (MLC), and a *Visiting Professor in Educational Psychology* (School Psychology) at the University of Minnesota.

Dr. McGrew authored the current document in his role as the Director of IAP. The opinions and statements included in this report do not reflect or represent the opinions of WMF, MLC, or the University of Minnesota.

More complete professional information, including his professional resume, can be found at www.iapsych.com.

Conflict of Interest Disclosure: Dr. McGrew is a co-author (with a financial interest) in the *Woodcock-Johnson Battery—Third Edition (WJ III; 2001)* as well as the *Batería III Woodcock-Muñoz (BAT III, 2005),* published by *Riverside Publishing*. He was a paid consultant, but was not a co-author, for the *Woodcock-Johnson Psychoeducational Battery—Revised* (WJ-R; 1989).