APPLIED PSYCHOMETRICS 101:

#13: Problems with the 1960 & 1986 Stanford-Binet IQ Scores in Atkins MR/ID Death Penalty Cases

Often in *Atkins* MR/ID death penalty cases historical and contemporary IQ scores are available for review by psychological experts. In many cases these scores vary markedly. The courts frequently wrestle with the issue of determining what the best estimate is of the person's general intelligence. A review of many *Atkins* cases often reveals frequent mention of two "*gold standard"* IQ tests in reports or testimony—namely, the *Stanford-Binet* and the *Wechsler* series.

The purpose of this working paper is to alert psychologists and the courts to two little known (but extremely important) dents in the gold standard status of two versions of the Stanford-Binet—the *1960 SB* and the *1986 SB IV*. If a Flynn effect adjustment is made to scores from a 1960 SB, the norm date used to calculate the magnitude of the Flynn effect should be 1932…not 1960. If SB IV scores exist in an individual's records, experts providing opinions regarding the individual's general level of intelligence should consider: (a) eliminating the score from consideration, (b) not give the score great weight in formulating an opinion, or (c) at a minimum, provide qualifying statements regarding the validity of the SB IV score as required by the *Joint Test Standards.*

Problems with the 1960 & 1986 Stanford-Binet IQ Scores

in Atkins MR/ID Death Penalty Cases

Working Paper: Kevin S. McGrew

(4-30-12 v1.0)

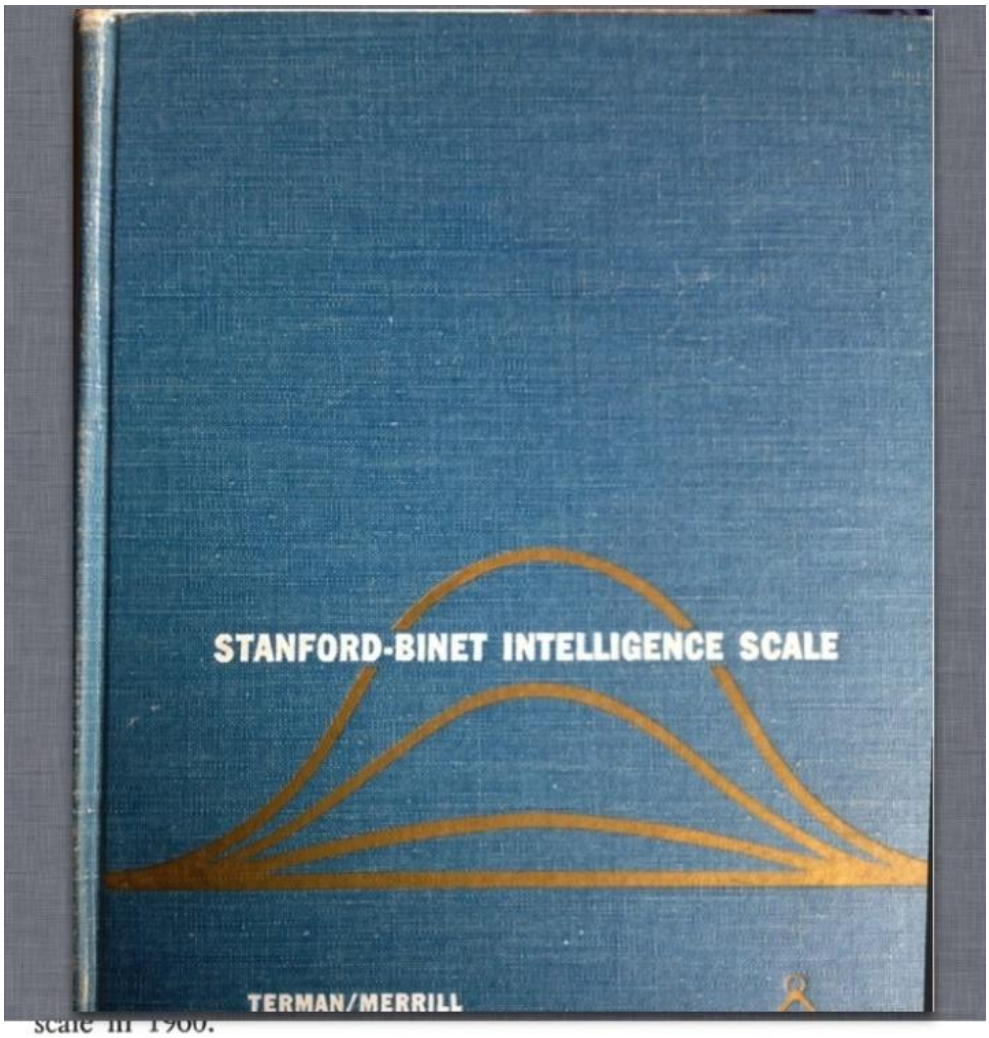
Often in *Atkins* cases historical and contemporary IQ scores are available for review by psychological experts. In many cases these scores vary markedly. The courts frequently wrestle with the issue of determining what the best estimate is of the person's general intelligence. Various experts weigh in using different methods. Because one of the three prongs of the AAIDD MR/ID definition requires evidence of MR/ID before the age of 18, any IQ test scores available from the person's formative years are critical.

However, many complex issues are encountered when examining historical pre-18 IQ scores, most notable of which is *norm obsolescence* (i.e., the *Flynn effect*). A review of many *Atkins* cases often reveals frequent mention of two "*gold standard"* IQ tests in reports or testimony—namely, the *Stanford-Binet* and the *Wechsler* series.

The purpose of this *IAP 101 Psychometric Report* is to alert psychologists and the courts to two little known (but extremely important) dents in the gold standard status of two versions of the Stanford-Binet—the *1960 SB* and the *1984 SB IV*.

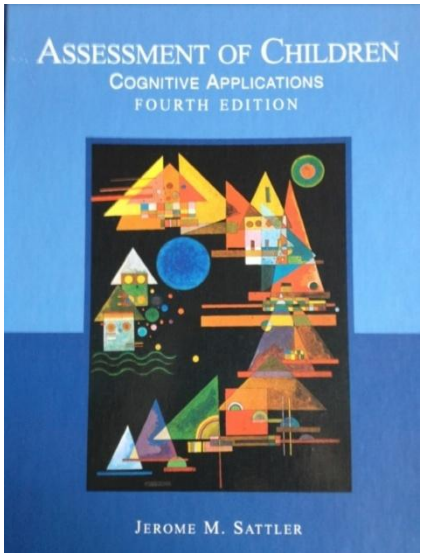
I. The 1960 SB test was not renormed and a large potential Flynn effect is not often recognized

Most psychologists assume that whenever a new revision of a test is published it also included the gathering of new nationally representative norm data. *This did not occur for the 1960 SB.* Below are excerpts from the 1960 SB manual and additional supporting excerpts from Sattler (2001). Although a sample of almost 5,000 individuals was tested, this was *not a nationally representative sample* and the data were used only for item difficulty verification and selection. *No new norms were calculated*. It was not until 1973 that a new standardization occurred for the SB.



scale in 1960.

We have been at some pains to review the well-established characteristics of Binet type tests. The scale in 1960 retains these major characteristics. Because of practical considerations we have not undertaken to restandardize the scales, but have undertaken to check existing standards against current empirical data. In the 1930's, for example, 69 per cent of the three-year-olds of the standardization group recognized and could name 5 out of 6 items consisting of miniature object reproductions of *shoe, watch, telephone, flag, jack-knife,* and *stove.* In the 1950's only 11 per cent of children



p. 459

1960 Stanford-Binet Scale: Form L-M

In 1960, Terman and Merrill revised the 1937 scales, although the 1960 edition was not a genuine revision. Terman and Merrill selected the best items from Form L and Form M of the 1937 scales and combined them into a new form. They did not obtain a new standardization group; instead, to check on changes in item difficulty, they used a sample of 4,498 participants who had taken the scale between 1950 and 1954. New material was not introduced, nor were the essential features of the scale changed. With the 1960 revision, there was only one form. The 1960 Form L-M manual did not present validity data; rather, validity rested on the fact that the same types of items were used as in the 1937 scales.

This fact means that anyone tested with the 1960 SB had their performance compared against the norm sample collected for the 1937 SB, a sample of subjects tested in 1932. With the 1973 norms collected from 1971 to 1972, this means that anyone tested with the 1960 SB in 1972 (the most extreme example) had their score based on norms that were 40 years old! With a Flynn effect rule-of-thumb of .3 points per year, this translates to an extreme Flynn effect of 12 points for a 1972 administration of the 1960 SB. This is the exact value that Flynn (1984) reported for the SB for the same time frame.

Conclusion and recommendation

The 1960 edition of the SB, which was in use from 1960 to 1972, produced scores based on a norm sample from 1932. Psychologists unaware of this fact run the risk of misinterpreting the meaning of 1960 SB scores. If a Flynn effect adjustment is made

to scores from a 1960 SB, the norm date used to calculate the magnitude of the Flynn effect *should be 1932…not 1960*. It is highly probable that the 1960 SB test was consistently reporting inflated IQ scores due to a very large norm obsolescence effect. Hopefully this little known (or long forgotten) fact will now be recognized and psychologists examining historical IQ records will recognize the need to apply a proper Flynn effect adjustment—not an erroneous one based on the assumption that the 1960 SB had new norms.

II. Psychometric Problems with the Stanford Binet IV (SB IV) IQ Test Battery

Soon after its publication in 1986, independent reviews and re-analysis of the SB IV standardization data, including studies and writings by some of the SB IV coauthors, raised serious questions about the reliability and validity of the SB IV Composite IQ score. Some of the most prominent evaluations and reviews reported the following technical problems with the SB IV.

1. Unrepresentative standardization or norm sample

The manner in which the SB IV standardization data was gathered was atypical and at variance with established test norm procedures and industry standards. The *SB IV final SB IV norm sample was deemed substandard and not representative of the US population*. Representative criticisms of the SB IV norming are presented below (emphasis via underling added):

- Walker (1987) stated: "Preliminary information from the *Technical Manual* suggests that the S-B IV standardization sample approximated the national population demographics with two noteworthy exceptions. Educational levels of the standardization sample seem to be substantially higher than national figures, with the S-B IV sample containing almost twice as many college graduates. Occupational characteristics were also somewhat inconsistent: The S-B IV sample contains a much higher percentage of managerial or professional respondents, obviously because of the high percentage of college graduates. The reason these discrepancies occurred is unclear. Although the results strongly suggest that the sample selection process may have been flawed (p. 137).

- Reynolds (1987) stated: "The stratification of the standardization sample, however, was far below industry standards. The preceding reviewer puzzled over why the S-B IV standardization sample differs so much from the U.S. Census Bureau data regarding the socioeconomic or educational level of the population. The answer is simple. Although the major publishers of individually administered tests (such as the Wechsler Scales and the Kaufman Assessment Battery for Children) routinely stratify samples by age, sex, race, geographic region,

> community size, and socioeconomic status, Riverside, declined to do so for the S-B IV. When I asked why, I was told that it was "too difficult" to stratify by so many variables. *Au contraire.* Not only is it not too difficult, it is the industry standard and is necessary for precisely the reason demonstrated in the S-B IV sample data. The sample is not adequately representative of the population of the United States, showing nearly twice as many individuals from the highest SES level as should be presented. Not noted in the preceding review is the fact that Riverside did weight the normative data in an attempt to mimic U.S. Census statistics and was moderately successful in achieving a fit (although calculation of chi-square shows that a significant difference, $p < .05$, remains). The weighting procedure, although necessary, does have a detrimental side effect. The confidence that typically would be engendered by such a large standardization sample (5,013 participants through age 23) in the estimation of the raw score distribution (the basis for the scaled scores) must be tempered considerably because the weighting procedure will increase the standard error of the mean of the raw scores. The sampling problems evident in the S-B IV add much error to the estimation of IQs (p. 140)

Test *norms* and *norm-referenced* test interpretation are defined by the *Joint Test Standards* in the following manner:

- *Norms:* Statistics or tabular data that summarize the distribution of test performance for one or more specified groups, such as test takers of various ages or grades. Norms are usually designed to represent some larger population, such as test takers throughout the country. The group of examinees represented by the norms is referred to as the *reference population* (p. 178) (italics in original).

- *Norm-referenced test interpretation:* A score interpretation based on a comparison of a test taker's performance to the performance of other people in a specified *reference population* (p. 178) (italics in original).

Norm-referenced testing is at the heart of psychological assessment for the diagnosis of MR/ID (AAIDD, 2010). The diagnosis of MR/ID requires comparison of a person's scores against *nationally representative norms.* This flaw in the SB IV norms represented a significant blow for use of the SB IV to establish, with a high degree of scientific confidence, the accuracy of a person's general level of intellectual functioning for the diagnosis of MR/ID.

2. Lack of comparability of IQ scores across the age levels of the SB IV

The SB-IV was constructed in an unusual manner that resulted in the total Composite IQ score being based on different combinations of tests at different age levels. The rationale for this variability was to provide for greater examiner "flexibility." Unfortunately, independent reviewers, and one of the SB IV co-authors, subsequently indicated that this was a significant flaw as it may cause a person's IQ score to vary depending on the age level of the subject and the decisions of the examiner. As described by Reynolds (1987),

this resulted in the potentially dangerous practice of "*IQ roulette."* Representative criticisms of the variable composition of the SB IV IQ scores are presented below (emphasis via underling and bold font added):

Walker (1987) stated:

- An advertised major advantage in S-B IV test administration procedures is the flexibility the examiner has in 'creating' appropriate batteries from among the remaining 14 tests for individual testing situations. The *Guide* recommends choosing batteries based on age (not all 15 tests have norms that were established on the full 2- year to adult age range), intent of the testing, time available, and expertise of the examiner. Because projected administration time for the entire 15-test battery is approximately 4 hours, it can be to the examiner's advantage to choose a more time-appropriate battery. An abbreviated, screening-only battery (Vocabulary, Bead Memory, Quantitative, and Pattern Analysis) purportedly requires 30-40 minutes to administer. Unfortunately, very little specific instruction is given on how to choose appropriate batteries in the *Guide.*

- A recommended "core" battery consists of only the six tests that cover the full normative age range (Vocabulary, Bead Memory, Quantitative, Memory for Sentences, Pattern Analysis, and Comprehension). Beyond this, however, the rationale for the six tests the authors suggest is unclear. Preliminary factor analysis, in fact, suggests that these six tests may not be appropriate for all ages, because they do not seem to adhere to the proposed hierarchical model (p. 136).

- Because there is no mandated "core" battery, examiners' composite SASs could be predicated on a seemingly endless variety of test battery combinations (p. 137).

Reynolds (1987) stated:

- The IQs and composite scores derived for the S-B IV are derived in an unacceptable manner (p. 140).
- The practical problems created by such a procedure seem insurmountable and make deriving IQs for anyone on the S-B IV, whenever less than all age-appropriate subtests are given, nothing short of IQ roulette. If you do not like the IQ you get from the S-B IV, change which subtests you give. The procedure of allowing examiners to pick and choose which subtests to give was a superlative idea, but norm tables must be available for every combination and could have been via a computer disk version of the norm tables. Without such tables, I believe the use of S-B IV IQs to be logically indefensible, and **I certainly would not want to have to defend their accuracy or validity in a court of law** (p. 141).

Sattler (1992), who was a coauthor of the SB IV, acknowledged significant problems with the SB IV's lack of comparability of IQ scores across ages:

> The SB: FE fails to provide a comparable battery of subtests throughout the age ranges covered by the scale. This is a limitation because it means that scores obtained by children at different ages are based on different combinations of subtests (p. 289).
>
> It has serious short comings, however, that must be recognized. The most serious of these are the lack of a comparable battery through the age levels covered by the scale and the nonuniformity of Composite Scores, factor scores, and scaled scores. Users of the scale must be extremely alert to these features of the scale (p. 290).

3. Complexity in administrating, scoring and interpreting the SB IV

The complexity involved in the administration and scoring of the SB IV was more than most other intelligence tests and increased the probability of administration and scoring errors. It also made competent interpretation of the SB IV difficult. A national survey of school psychologists (Chattin & Bracken, 1989) evaluated their ratings of four major intelligence batteries commonly used at the time (K-ABC, McCarthy Scales, SB IV, WISC-R). Select highlights related to the SB IV are reported below (italics in original; emphasis added via underline):

> The WISC-R was viewed as the easiest test to administer, and the SBIV was the most difficult (p. 116).
>
> The majority of the respondents believed that they had received adequate training on all of the tests except the SBIV (p. 116).
>
> *Manipulating materials* on the SBIV was rated as the overall most difficult of the seven areas across the four tests (p. 117).
>
> The SBIV fairly consistently received the lowest ratings throughout the survey. The test is viewed as significantly more difficult to administer and interpret than any of the remaining tests (p. 126).
>
> It appears that practitioners want instruments that, while not simple, are theoretically parsimonious. They also seem to want tests that are conveniently administered, practically interpreted, and yield meaningful information. With regard to these attributes, the SBIV was a major disappointment. The theoretical factor mismatch renders interpretation somewhat bothersome (McCallum, in press; Sattler, 1988), and the delay in publication of the administration, interpretation, and technical manuals resulted in a useless test, long after the test materials were shipped by the publisher (p. 128).

4. Variable Findings in Concurrent Validity Studies

A series of independent articles reported that the SB IV IQ scores were significantly higher than the WISC-R IQ scores in samples of children with MR/ID or, the SB IV/WISC-R IQ differences varied as a function of the level of intelligence of the children with MR/ID (Lukens, 1990; Prewett & Matavich, 1992; Prewett & Matavich, 1994). These independent studies were at variance from the SB IV/WISC-R IQ score comparisons for children with MR/ID as reported in the SB IV Technical Manual (Thorndike, Hagen & Sattler, 1986).

Additional non-systematic WISC-R/SB IV studies were also reported for groups of gifted children and children with learning disabilities. In a comprehensive synthesis and review of the available SB IV research, Laurent, Swerdlik and Ryburn (1999) reported that there were "wide variations when examining the validity coefficients for groups of exceptional children" (p. 106). It is important to note that Laurent et al. (1999) concluded that the SB IV "appears to be as good a measure of *g* as other existing measures of intelligence, especially for nonexceptional populations" (p. 108; emphasis via underlining added). A nonexceptional population refers to individuals typically within the normal range of intelligence and without disabilities.

When a valid IQ test battery is compared to other IQ test batteries, well-designed concurrent validity studies are expected to show systematic relations between the IQ scores obtained on the respective instruments in the form of high correlations. In addition, if average mean score differences are found (i.e., Test A scores higher than Test B), if an IQ battery is validly measuring the construct of general intelligence, these mean score differences should be in a systematic direction across groups with exceptionalities or disabilities where general intelligence plays a prominent role in defining the condition (e.g., MR/ID, learning disabled, gifted). The extant SB IV/Wechsler IQ comparison research, although often reporting significant correlations, demonstrated a non-systematic trend for SB IV scores when compared to other established and validated IQ batteries.

Although other criticisms were also reported (e.g., inconsistent factor structure; Laurent et al., 1999), the above reviews of the SB IV and independent research suggested significant flaws in the psychometric foundations of the SB IV. Collectively, the scientific evidence accumulated regarding the SB IV, as well as independent reviews and post-publication comments by some SB IV co-authors, led to the conclusion that the SB IV was a sub-standard IQ test battery that provided IQ scores of questionable reliability and validity. In addition, it was prone to examiner administration and scoring errors. It was also difficult to interpret. The professional consensus regarding the SB IV is best captured by the following summary comments:

Walker (1987).

> One of the most disappointing features of the S-B IV is the almost reckless haste with which it was made available to the public (p. 138).
>
> Much "rough edge rounding" needs to be done, however, before it should be put into use. As of this writing two states (California and Florida) have voiced serious reservations about the S-B IV with regard to its use in placement decisions for exceptional programs in public schools. A third (North Carolina) seems to be following suit (p. 138).

Reynolds (1987).

> As the other reviewer in this issue has noted, the questionable quality of the revision results in the unfortunate situation that the name Binet has become associated with an intelligence scale of questionable quality and utility (p. 139).
>
> The S-B IV was developed with a set of noble goals and intentions and is the broadest, most flexible, and most versatile individually administered intelligence test available. This is no small claim and could have led the S-B IV to a clearly dominant position in the field. Its failure to do so can be attributed almost entirely to the inadequate technical execution of the rich ideas of the test's authors (p. 139).
>
> The S-B IV revision was a spectacular conception. However, it was executed, particularly in the presentation in the *Technical Manual* with much psychometric naïveté and a lack of savvy regarding individually administered tests. Riverside does an excellent job of publishing group tests. Perhaps their experience with the S-B IV will convince them to stay in that domain. To the S-B IV, *Requiescat in pace:* and so it should have stayed (p. 141).

Although the old (prior to the SB IV) Stanford-Binet IQ test is often described as one of the two "gold standard" IQ batteries in *Atkins* psychological reports and testimony, and the more recent SB V is worthy of this characterization, the "gold standard" characterization no longer held true in the case of the SB IV. Much like Toyota once was considered a gold standard for automobile safety (a distinction which it recently lost as a result of the documented safety problems with the Toyota Camry) the SB IV was a Toyota Camry-like drop from gold standard status for the venerable Stanford-Binet.

Conclusion and recommendations regarding the use of SB IV IQ scores in *Atkins* cases

Soon after the SB IV was published independent reviews and research established that the SB IV was extremely hard to administer and was thus prone to administration and scoring errors. It was also shown to be hard to interpret. It was heavily criticized on the ground that it could produce different IQ's for the same individual based on varying subtest selection decisions by individual examiners (i.e., *"IQ roulette"* criticism of SB IV). More importantly, the scientific evidence and

professional consensus was that the SB IV standardization norm sample was flawed and most likely failed to meet the minimum test development industry standards as established by the *Joint Test Standards*.

It is recommended that SB IV IQ scores not be given great weight in the determination of a person's general intelligence, particularly when making high stakes decisions. If SB IV scores exist in an individual's records, experts providing opinions regarding the individual's general level of intelligence should consider: (a) eliminating the score from consideration, (b) not give the score great weight in formulating an opinion, or (c) at a minimum, provide qualifying statements regarding the validity of the SB IV score as required by the *Joint Test Standards.*

References

American Association on Intellectual and Developmental Disabilities. (2010). *Intellectual disability: Definition, classification, and systems of supports—11th Edition*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Chattin, S. H. & Bracken, B. A. (1989). School psychologist's evaluation of the K-ABC, McCarthy Scales, Stanford-Binet IV, and WISC-R. *Journal of Psychoeducational Assessment, 7,* 112-130.

Choi, H-S. & Proctor, T. B. (1994). Error prone subtests and error types in the administration of the Stanford-Binet Intelligence Scale: Fourth Edition. *Journal of Psychoeducational Assessment, 12*, 165-171.

Flynn, J. R. (1984a). IQ gains and the Binet decrements. *Journal of Educational Measurement, 21*, 283-290.

Laurent, J., Swerdlik, M., & Ryburn, M. (1992). Review of the validity on the Stanford-Binet Intelligence Scale: Fourth Edition. *Psychological Assessment, 4(1),* 102-112.

Lukens, J. (1990). Stanford-Binet, Fourth Edition and the WISC-R for children in the lower range of intelligence. *Perceptual and Motor Skills, 70*, 819-822.

Prewett, P. N. & Matavich, M. A. (1992). Mean-score differences between the WISC-R and the Stanford-Binet Intelligence Scale: Fourth Edition. *Diagnostique, 17(3),* 195-201

Prewett, P. N. & Matavich, M. A. (1994). A comparison of referred students' performance on the WISC-III and the Stanford-Binet Intelligence Scale: Fourth Edition. *Journal of Psychoeducational Assessment, 12,* 42-48.

Reynolds, C., Niland, J., Wright, J., & Rosenn, M. (2010). Failure to Apply the Flynn Correction in Death Penalty Litigation: Standard Practice of Today Maybe, but Certainly Malpractice of Tomorrow. *Journal of Psychoeducational Assessment, 28 (5),* 477-481.

Reynolds, C. R. (1987). Playing IQ roulette with the Stanford-Binet, 4th Edition. *Measurement and Evaluation in Counseling and Development, 20 (3),* 139-141.

Sattler, J. (2001). *Assessment of Children: Cognitive Applications—4th Edition*. San Diego, CA: Jerome M. Sattler, Publisher, Inc.

31, 499-506.

Thorndike, R. L., Hagen, E. P. & Sattler, J. M. (1986). *The Stanford-Binet Intelligence Scale: Fourth Edition, technical manual*. Chicago: Riverside.

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 14 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. Dr. McGrew is a coauthor of the WJ III battery. The opinions and statements included in this report do not reflect or represent the opinions of WMF, MLC, or the University of Minnesota. The opinions and statements included in this document do not necessarily reflect the opinions of the publisher of the WJ III Battery (Riverside Publishing) or the other WJ III co-authors.

More complete professional information, including Dr. McGrew's professional resume, bio, and conflict of interest disclosures can be found at each of his three professional blogs and web page:

- www.iqscorner.com
- www.atkinsmrdeathpenalty.com
- www.ticktockbraintalk.blogspot.com
- www.themindhub.com