IAP AP 101

APPLIED PSYCHOMETRICS 101:

#5: The Standard Error of Measurement (SEM): An Explanation and Facts for "Fact Finders" in Atkins MR/ID death penalty proceedings

The standard error of measurement (SEM) is a professionally accepted and scientifically based measurement concept that allows users of psychological test scores to account for the known degree of imprecision in the scores. Atkins MR/ID cases almost always involve standardized psychological testing in the domains of intelligence (IQ tests) and adaptive behavior (AB). Scores from IQ and AB measures are fallible—not perfectly reliable. This report provides an easy to understand explanation of the psychometric concept of SEM augmented by an example based on real-world data. The report concludes with 8 SEM facts that "fact finders" should understand and internalize when evaluating psychological test data during legal proceedings--Atkins MR/ID death penalty proceedings in particular.

Kevin S. McGrew, Ph.D.

Educational Psychologist

Director

Institute for Applied Psychometrics (IAP)

> **Standard error of measurement:** The variation around a hypothetical "true score" for the person. The standard error of measurement apples only to scores obtained from a standardized test and can be estimated from the standard deviation of the test and a measure of the test's reliability. The standard error of measurement, which varies by test, subgroup, and age group, should be used to establish a statistical confidence interval within which the person's true score falls…Reporting the range within which the person's true score falls, rather than only a score, underlies both the appropriate use of intellectual and adaptive behavior assessment instruments and best diagnostic practices in the field of ID. *Such reporting **must** be a part of any decision concerning the diagnosis of ID* (italic and bold font emphasis added by author) (AAIDD, 2010, p. 224)

Introduction to Measurement Error and the Standard Error of Measurement

Four times a year the future lawyers and judges of our country anxiously await the receipt of their LSAT (Law School Admission Test)[1] examination results. Given the prominent role LSAT scaled scores (scores range from 120-180, average of approximately 151) [2] play in ABA-approved law school admission procedures, the future of many prospective law school students may hinge on the specific LSAT score they obtain. To many applicants, their LSAT score, particularly if precariously close to the specific cut-score of their law school of choice, is anxiously perceived as dictating a life-or-death decision regarding their chosen law career path. Fortunately for potential students near the cut-points, the LSAC recognizes that no standardized test (including the LSAT) is perfectly reliable and that *measurement error* must be factored into the interpretation of an individual's score. According to the Law School Admission Council:

> The LSAT, like any standardized test, is not a perfect measuring instrument. One way to quantify the amount of measurement error associated with LSAT scores is through the calculation of the **standard error of measurement**. The standard error of measurement provides an estimate of the average error that is present in test scores because of the imperfect nature of the test. An error-free score, called a true score, could only be obtained from a hypothetical test that contained no measurement error. This brochure explains score bands, which are used in score reports to quantify the uncertainty inherent in individual test scores. Many factors besides measurement error can also affect an individual's test performance on a particular day (e.g., motivation, physical and mental health, or work and family responsibilities). These other factors are not explicitly taken into consideration when calculating score bands.

For a young person who dreams of attending Harvard Law School, a score of approximately 170 or above is desirable.[3] But what if this student receives an LSAT score of 169? Should this score, which is one point below the 25th percentile of Harvard Law School applicants, mean plans for a Harvard law education are terminated immediately? Fortunately, the Law School Admission Council (LSAC) incorporates the accepted professional and scientific psychometric practice of recognizing that any LSAT score contains a known degree of measurement error. More importantly, LSAC recognizes that each specific LSAT score should be surrounded by a confidence band (a range of scores that reflects known measurement error). LSAT scores have a standard error of measurement (SEM) of 2.6 points (we round to 3 for ease of presentation) (Law School Admission Council, 1997), As will be explained later, the incorporation of the LSAT SEM-based confidence band allows he Harvard admission "deciders" to be 68% confident (1 SEM) that the student's "true" LSAT score would be somewhere between 166 (-1 SEM or -3 points) and 172 (+1

[1] The LSAT is administered by the *Law School Admission Council (LSAC)* and is designed to assess logical and verbal reasoning skills.

[2] http://www.lsatprepcourse.com/law.htm

[3] The average (between 25th and 75th percentile ranks) Harvard LSAT score is 170-175.

SEM or +3 points). They can be 95% confident that this student's true LSAT score is in the range of 163-175 (± 2 SEM or ± 6 scaled score points). Clearly the admission committee, when giving weight to the LSAT score (along with other information and/or scores), should recognize that falling one point below the bright line cut-score of 170 [4] should not result in the "death penalty" for this student's possible admission to Harvard Law School. The SEM confidence band is necessary to insure that the known degree of imprecision (lack of perfect reliability) of the LSAT is factored into this critical life event decision for each and every applicant.

If LSAT scores for individuals being considered for law school admission are to be interpreted within the context of the LSAT's known degree of measurement error (as reflected by the SEM), shouldn't it follow that the IQ scores for individuals being considered for a diagnosis of *mental retardation(MR)* or *intellectual disability (ID)*[5] (and potential capital punishment) be accorded the same equal protection as per the professionally and scientifically grounded constitution of psychological measurement (viz., the Joint Test Standards)?[6]

My answer to this question is obvious—yes. Unfortunately, a portion of the misunderstanding and misapplication of the IQ SEM rule in MR/ID diagnosis may be due to the construct and statistically dense measurement and statistical concepts often used by psychometric specialists to explain reliability and SEM. Hypothetical true score. Observed or measured score. Reliability. Different types of reliabilities. Standard deviation. Standard error. SEM. Conditional SEM. SEM-based confidence interval around the observed score versus SEM-based confidence interval around the true score. Symmetrical versus asymmetrical confidence bands. 68 %, 90%, 95% confidence interval---which to use? Ability-centered, local, conditional, or overall reliability and SEM. $SEM = SD * \sqrt{1 - r(11)}$.

The dizzying array of quantoid-speak and statistical symbols can understandably result in confusion and frustration. In reality, the concept of SEM is very simple when one sets aside the psychometric "slicing of the measurement and statistical butter with a laser beam" discourse that occurs when scholarly psychometricians discuss and debate the subtle nuances of SEM, nuances that are not necessary for practical implementation of SEM-based IQ decisions by the courts.[7] The goal of this IAP AP101 report is

---

[4] For illustrative purposes I have arbitrary assumed that the Harvard Law School admission process has a bright-line or cut-score of 170. This is a hypothetical bright line and in no way reflects the use of specific scores as per existing Harvard Law School admissions policies.

[5] The field of mental retardation, as led by the American Association on Intellectual and Developmental Disabilities (AAIDD), now refers to mental retardation as an *intellectual disability (ID)*. Unfortunately, currently (and most likely for a number of years to come), the courts continue to use the older term mental retardation (MR). Because of this terminology "bridging" time, the abbreviation MR/ID will be used throughout the rest of this paper. According to AAIDD (2010), an intellectual disability (ID) is *"a disability characterized by significant limitations in both intellectual functioning and in adaptive behavior as expressed in conceptual, social and practical adaptive skills. This disability originates before age 18"* (p. 221).

[6] The *Standards for Educational and Psychological Testing* (1999), known as the "Joint Test Standards," are the professional gold standards for those who develop, publish, and use psychological and educational tests. The standards are a collaborative effort of the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The current version is published by AERA.

[7] The reader interested in the various nuances of reliability theory and the standard error of measurement is urged to read two relatively concise instructional modules produced by the *National Council on Measurement in Education* (NCME). The module *Understanding Reliability* (Traub, 1991) is available for download. The *Standard Error of Measurement* module (Harvill (1991) is available for download. The complete set of NCME instructional modules that address an array of psychometric topics can be found at an NCME web page. Although there are some technical differences in how reliabilities and SEM's may be calculated, the inclusion of these technical nuances would make the current explanation unnecessarily detailed and complex. More importantly, these technical issues, which are important to psychometricians and psychometric theory, are mute points in the current context given that all major individualized IQ tests typically use a common set of statistical procedures for calculating the reliability of the total or full scale IQ score and its SEM.

to rectify this problem via the explanation of SEM in understandable English augmented by real-world data.

Some Basic Standardized Testing Concepts

Anyone who has regularly played a game or performed in a domain judged via objective scoring criteria (e.g., bowling, bridge, chess, running marathons) is personally aware of the variability in their performance from day-to-day, game-to-game, or event-to-event. No one, not even the most accomplished expert or star within a performance domain, operates at optimal level on all occasions. For example, on any given day an individual may suffer from fluctuations in levels of energy, mental concentration, and physical well-being, to mention but a few within-person variables. External factors beyond the individual's control (e.g., refurbishing the bowling alley lanes between tournaments; environmental temperature being uncharacteristically hot or cold, etc.) may also influence performance. However, over many different occasions a general consistent pattern of typical or average level of performance emerges (e.g., bowling average; golf handicap), together with a range of typical variability in performance for each individual. The same holds true for performance on individually administered IQ tests. The "quantification of the consistency and inconsistency in examinee performance constitutes the essence of reliability analysis" (Feldt & Brennan, 1989, p. 105), and the important statistical estimate derived from a test's reliability—the standard error of measurement (aka., SEM or "standard error").

The quantification and objective comparison of different individuals in human performance situations is made possible via the use of standard sets of tasks and measurement procedures. For example, when measuring bowling performance, each "test" session is the same via the holding constant of the number of frames bowled, length and width of the bowling lanes, size and weight of the bowling pins, and the use of a 100% (barring cheating or addition errors) objective scoring method. As a result, each bowling session is made as identical as possible to all other sessions to allow each person's bowling ability to be judged via an objective common yardstick. The key to precise, dependable, and consistent measurement in any human performance domain is *standardization*. So it is with standardized IQ testing.

"A standardized test is a task or set of tasks given under standard conditions and designed to assess some aspect of a person's knowledge, skill, behavior, or personality" (Green, 1981, p. 1001). Individually administered standardized IQ tests are designed to reduce sources of error in assessment that may result from idiosyncratic or biased assessment methods used by any individual psychological examiner. This is achieved via the use of standard or uniform procedures for (a) test item content administered to examinees, (b) prescribed test administration procedures (e.g., wording of items and directions; time limits; etc.), and (c) objective scoring criteria (Sattler, 2001). The use of standardized testing procedures insures "a constant testing environment and conducting the test according to detailed rules and specifications, so that testing conditions are the same for all test takers" (AERA, APA, NCME, 1999, p. 182). Thus, the result of a standardized IQ test is an objective quantifiable score that is purged, to the maximum degree possible, of idiosyncratic characteristics of the person administering the test. The result is a fair, equitable and standard comparison of examinees measured via a common method or mental yardstick (Green, 1981).

Just as perfectly standardized measurement conditions prove elusive in all human performance situations (e.g., despite a scoring system with perfect reliability, in a golf tournament one set of golfers starts early enough to enjoy perfect weather conditions while players starting later in the day may face more adverse weather conditions), standardized IQ tests are fallible or imperfect measures of an individual's level of general intellectual ability. However, due to the development of the psychological specialty of psychometrics, a well established, professionally accepted, mathematically tractable statistic (standard error of measurement or SEM) has been developed to quantify the amount of potential error in any IQ score.

Returning to the sports examples described above, across multiple standardized performance situations common in organized sports, an estimate of a person's "true" bowling, golfing, bridge, or marathon ability (score) eventually emerges. This true ability score or index is reflected in the person's average

performance across all measured performance situations. And, of course, there is a range within which the person typically scores or performs "most of the time." The psychometric idea of a person's "true IQ" or "true ability" is based on the same concept. If we were to administer a person the same IQ test over and over again (e.g., 100 or more times), at the end of this process a range of typical IQ scores, as well as the arithmetic average across all 100 IQ test scores, would emerge in a manner similar to a professional bowlers average or a person's PGA golf handicap. The IQ scores would not be identical every time due to temporary fluctuations and chance conditions present during different testing sessions. However, as in most human performance situations, the person's most typical scores would tend to cluster or group together in the middle of the complete range of the person's high and low scores, with the frequency of extremely high or low scores decreasing the farther away the extreme scores are from the middle of the center of the score grouping. The average of repeated measurement with the same IQ test is what measurement experts refer to as the best estimate of the person's "*true ability*" or "*true IQ score*."

But wait a minute. The norming of standardized IQ tests does not involve the repeated testing of all norm subjects in order to obtain every subject's true IQ score. IQ testing is not like golf or bowling where performers rack up hundreds and hundreds of scores for analysis. In IQ test development, each norm subject is typically tested only once. Also, no person is ever administered the same IQ test 100 or more times in clinical or forensic settings in the hunt for his or her true IQ score. Thus, a person's true IQ score is a figment of measurement expert's imaginations (Kaufman, 2009). It is never possible (nor would it be humane or ethical) to administer the same IQ test 100 times to one individual. A person's true IQ score is a hypothetical concept. We will never know a person's true IQ score. However, all is not lost. The field of psychometrics, via the use of standardized testing procedures and the development of methods to calculate a test's reliability (the precision, consistency, and repeatability of the test score), has provided a mathematical means to estimate the amount of known variability or measurement variability that would be present if a person could be tested 100, 200, or more times. Once an IQ test's reliability is known, a simple mathematical calculation can produce an estimate of the average variability of the observed or measured IQ scores expected across all persons taking the test. This is the standard error of measurement (SEM).

A Real-World Example of SEM

Rather than engaging in quantoid-speak and the temptation to demonstrate expertise in statistical symbol manipulation and equation wizardry (and bore the reader to sleep) in this report, a real-world example of the measurement concepts of reliability and SEM is presented instead. The example involves a young male who, over a period of 13 years, had his ability tested 52 times with the same standardized procedures. The distribution of his obtained or measured scores (not his true ability scores) is presented in Figure 1. [Note..the quality of the image in the final PDF file is not optimal. A higher quality copy of Figure one can be obtained, and saved, by clicking here.]

Most readers will immediately be struck by the approximate shape of the histogram graph, where each column represents the frequency of each measured ability score. Yes. It is a visual-graphic that may strike fear into the psyches of many who, during their professional education, took one or more courses in statistics—the normal distribution or bell curve.[8] Although the observed or measured scores graphed in Figure 1 do not represent a perfect normal distribution, if this unique individual was tested another 50+ times the shape of the

[8] The bell shaped normal distribution in Figure 1 was calculated via standard statistical software that used the complete collection of ability scores portrayed in Figure 1.

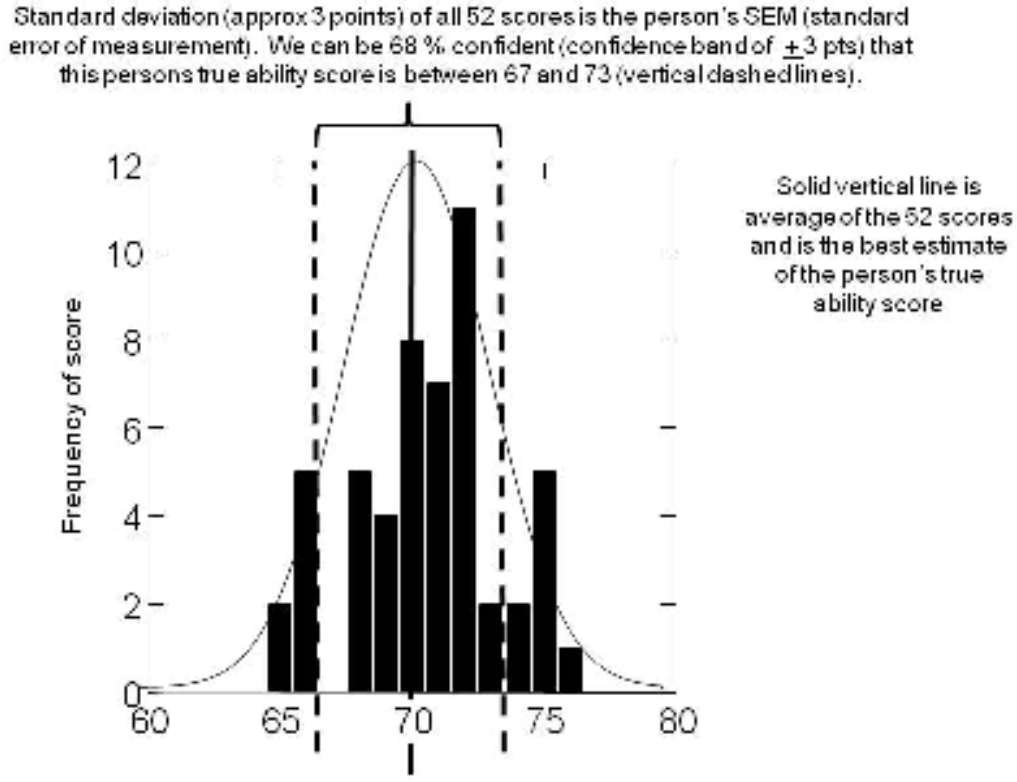

Figure 1. Distribution of 52 scores obtained by one individual across 13 years of standardized ability performance

histogram would, with each new set of scores, more closely approximate the normal curve. Even with this small set of scores (small in the world of statistical calculations), the reader should see (in Figure 1) the general clustering or grouping of most of the person's scores near the middle, and the more extreme high or low scores occurring less frequently. The normal curve shape presented in Figure 1 is the mathematical model underlying the distribution of this set of scores and has been found to be a valuable and reasonably accurate approximation of the distribution of many human traits (e.g., height, cognitive ability) and natural phenomena (Shavelson, 1996). The field of psychometrics has demonstrated, that if it were possible to test the same person with the same ability measure repeatedly, and if the measurement procedure is purged of measurement error as much as possible via standardized measurement procedures, fluctuations between observed scores will primarily be a function of random variables (error) which, because they operate randomly in either direction, will form a distribution that approximates the normal curve (as illustrated in Figure 1) (Kaufman, 2009).

Anytime observed data for a phenomenon approximates the normal curve, statisticians and psychometricians can apply a cadre of proven, tractable, and rigorous statistical tools to characterize the data. The calculation of the typical or average score (e.g., mean, median, mode) is possible. In Figure 1 the average (*mean*) score across all ability scores for this person is 70.6 (represented by the thick black vertical line)—rounded down to 70 for ease of calculations and presentation. As defined and described previously, the mean ability score of 70 is the best estimate of the individual's hypothetical true ability score. Another normal curve-based tool is the calculation of the *standard deviation (SD)* which is a numerical index quantifying the degree to which the scores deviate from the average or typical score (70 in the current example) in a distribution of scores. The SD in Figure 1 is represented by the two dashed vertical lines. The calculated SD for the observed scores in Figure 1 is 2.8 (rounded to 3.0 for ease in calculations and presentation). As most readers will recall from their statistics courses, the point from the -1 SD line (67) to the mean (70) accounts for approximately 34% of the person's observed scores while the point from the mean to +1 SD (73) accounts for a similar 34 % of the person's observed scores. Collectively, 68% of this person's observed scores fit between -1 SD (67) and +1 SD (73). In the context of the current example (one person's ability scores over 52 different occasions on the same standardized set of tasks), the SD represents the person's standard error of measurement (SEM) (Shavelson, 1996). Thus, if this person were to take the same standardized ability test again, we can be 68% confident that the person's true ability score would be somewhere in the range of 67 to 73 (Shavelson, 1996). This range of scores is known in statistical terms as the 68 % *confidence band or confidence interval* for the ability score (Sattler, 2001). Via the known mathematical characteristics of the normal curve, we can be 95% confident that the person's true ability score will fall from -2.0 to +2.0 SEM units (2 x 3.0 =6.0) above and below the average score of 70—a 95% confidence band from 64 to 76.

The only difference between the real-world individualized SEM in our example and that previously discussed for the LSAT, is that our current example is a *personalized* SEM—it is estimated from a single person's tested ability over multiple testing sessions. This is the ideal. However, as noted previously, IQ tests are standardized on the single administration of the test to a large number of individuals. It is not necessary to demonstrate the math here, but in the case of the standardized norm-referenced ability tests (e.g., LSAT, IQ tests), although we cannot obtain a personalized SEM for each norm subject, the beauty of large-scale numbers (thousands of subjects take the same IQ test when it is normed), when combined with tractable psychometric logic and statistics, is that we can determine, from the test's calculated reliability, the average SEM across individuals (Traub, 1991). This average SEM (e.g., 2.6 in the case of LSAT scores) is then applied to norm-referenced ability tests such as the LSAT and standardized IQ tests. In the case of the major individually administered standardized norm-referenced intelligence tests with adult norms (i.e., *Stanford-Binet—5th Edition*; *Wechsler Adult Intelligence Scale—IV*; *Woodcock-Johnson Battery III*), 1 SEM is remarkably consistent for each battery's total IQ score for the adult subject norms (1 SEM = 2.12, 2.16, 2.12, respectively). Given that it is professionally recommended best practice (when making MR/ID determination) to bracket an individual's observed or measured IQ with a 95 % confidence band (±2 SEM), a person's obtained IQ score would be bracketed with a range of scores extended downward 4-5 points (e.g., 2.12 x 2 = 4.24) and upward a similar 4-5 points. Typically in the MR/ID classification literature (AAIDD, 2010), and almost always when discussed in Atkins MR/ID

determination, the rule of thumb of ± 5 IQ points (for a confidence band that spans 10 IQ points) is recommended.

Before proceeding, a confession of deliberate (but educationally necessary) deceit must be admitted, deceit that may have been recognized by astute readers. No young male ever took the same IQ test 52 times. However, the data in Figure 1 are real performance data. The data are the scores for the last 52 rounds of golf Tiger Woods achieved on the same golf course (Augusta National Golf Club—home of the annual Masters golfing tournament).[9] Granting the author a few degrees of freedom in the use of this example, the 18-hole Masters course can conceptually be considered a standardized set of tasks, much like the standard set of subtests in an IQ battery. Tiger Woods (and other golfers) perform on the same 18 holes[10] each year with a rigid set of prescribed rules, directions, and an objective (perfectly reliable) scoring system. As can be seen in Figure 1, Tiger did not receive the identical golf ability score on each of the 52 rounds of golf. It flies in the face of logic and the laws of nature to expect Tiger Woods to obtain the identical golf score on each round at the Masters or, to always receive the best estimate of his true Masters ability score (70) on any given round. It is just as nonsensical to expect a person taking a standardized IQ test to obtain the exact same IQ score if tested more than once.

Tiger Woods' obtained golf scores allow us to harness the characteristics of the normal curve to provide an objective estimate of the variability of expected scores and the confidence we can place in expecting his true ability score to be within a certain range. Tiger's scores show a distinct pattern with the majority in the middle of the histogram (the clustering or grouping of scores) and the more extreme high and low scores occurring less frequently. One can state that the best estimate of Tiger Woods' true Masters golfing ability is 70. More importantly, it is recognized that individual (e.g., level of concentration; physical health) and external factors (e.g., different weather conditions; different pin placements) occurring during any given round (test) result in variability in Tiger's scores across repeated "testing." And, we can quantify his personal Masters standard error of measurement (SEM) as approximately ± 3 strokes or points, from which we can specify any % confidence band to characterize the known error of measurement of his golfing ability on this particular golf course (test). As demonstrated above, the SEM summarizes the within-person inconsistency for a person when tested via imperfect standardized measurement situations.

---

[9] The development of this report, inclusive of the Tiger Woods Masters golf tournament performance data, was initiated prior to recent personal revelations (2009) regarding Tiger Woods personal life and behavior. The use of Tiger Woods Master's performance data in no way reflects an endorsement or judgment of his personal behavior. Prior to this revelation, he was clearly the world's best golfer. His Master's performance data is used since most individuals, whether golfers or not, recognize his name and golf accomplishments.

[10] In reality the 18 golf holes vary from year-to-year and round-to-round via the varying of pin placements, removal of trees, changing the length of the rough, adding or changing sand traps, etc. Thus, golfing at the Masters is not a perfect analogy to standardized testing. However, one could consider these round-to-round or year-to-year variations to be similar to the known variability in test administration of IQ tests by different examiners, variability that introduces a degree of error into the measurement of intelligence.

It is interesting to examine each of Tiger Woods' Masters scores in the context of his personal Masters SEM. As reflected in Figure 1, Tiger Woods has obtained 10 different Masters golfing scores (every score from 65 to 76 except 67). Each score and resulting 95% confidence interval (2 SEM = ± 6) is listed below:

Score	95% Confidence Band
65	59 to 71
66	60 to 72
68	62 to 74
69	63 to 75
70	64 to 76
71	65 to 77
72	66 to 78
73	67 to 79
74	68 to 80
75	69 to 81
76	70 to 82

All of the 95% confidence bands include the best estimate of Tiger Woods' true Masters golfing ability (70), although the score of 70 is at the bottom limit of the confidence band surrounding his highest score of 76. Nevertheless, these scores and confidence bands demonstrate the extent to which SEM confidence bands provide a solid zone or range of potential scores that most likely includes the person's true ability score.

Sources of Measurement Error in IQ Testing

It is impossible to list all potential sources of measurement error that may be captured by the SEM during IQ testing. However, the sources are typically divided into two broad categories. The first is *random variation within or internal* to the individual during testing. These might include less than optimal health, fluctuations in mental efficiency or motivation, fatigue, ability to comprehend instructions, emotional state, forgetfulness, ability to ignore distractions, carelessness, impulsiveness in responding, and willingness to guess (and be lucky) in random guessing.[11] Tiger Woods' best score of 65 may have occurred during a Masters round when he was in top physical condition, had extraordinary concentration and focus, and his ball took some lucky (random) bounces on particular holes. In contrast, his highest (poorest) score of 76 may have occurred on a day of poor health, when recovering from physical injuries, or from unfortunate (random) bounces of his ball on the fairways of certain holes. Similar factors also may occur during individualized IQ testing.

The second class of potential sources of error in IQ scores come from sources *external to the person tested*. These are *environmental factors*. Illustrative (but not exhaustive) examples include the time of day of the assessment, location of the testing room near external noise, unexpected interruptions by individuals inadvertently opening the testing room door, poor room conditions (heating, lighting or ventilation), failure of the examiner to follow standardized directions on certain tasks, and scoring errors by the examiner, to name but a few. Although the Masters scoring system is most likely perfectly reliable (which is not the case in standardized IQ testing), differences between some of Tiger Woods' best and worst Masters golfing scores may have been due to significant differences in weather conditions, slightly different conditions of the grass and greens, and other environment variables that varied from one Masters round to another. So it is with standardized IQ testing. Factors external to a person being tested, which have nothing to do with the person's intellectual ability, may result in an obtained IQ score that over- or under-estimates the person's true IQ.

[11] These factors represent a synthesis of factors mentioned by Feldt & Brennan (1989) and Sattler (2001).

Summary of SEM for Legal Fact Finders

In criminal proceedings the person(s) responsible for determining which facts have been proven is called the "fact finder." In this concluding section on SEM, the following facts are provided to assist fact finders perform their legal duty:

1. It is a professionally accepted and scientific fact, derived from the field of psychometrics, that "if a person could be tested thousands of times, the test scores would form a normal curve. The mean of that normal distribution of scores is the person's true score. The standard deviation is the standard error measurement" (Kaufman, 2009, p. 141). The example of 52 rounds of golf by Tiger Woods on the Masters course (test) is a small sample demonstration of this fact.
2. It is a fact that one cannot administer the same IQ test to one person hundreds and hundreds of times in pursuit of the person's true IQ score. Instead, an IQ score reported from any single administration of an IQ test must be considered to be one estimate of the person's true IQ, and the SEM allows us to quantify the amount of confidence that the SEM-based range of scores includes the person's true IQ.
3. It is a fact that the field of psychometrics has harnessed the information from large-scale numbers (obtained during the national norming of a test) to capture the mathematical tools of the normal curve which, when combined with the ability to calculate an IQ test's reliability, provides a mathematically rigorous, precise and tractable set of calculations for determining the average SEM to expect across all individuals for an IQ test.
4. It is a fact that the above facts (1-3) are neither new nor unproven. These facts have been professionally accepted as scientifically sound for decades. The characteristics and application of the normal curve derived mathematical tools to the measurement of human traits have been known and established as early as 1786-1835 (Sattler, 2001).[12] It is also a fact that the mathematical assumptions, logic and method for calculating a confidence band (based on a test's SEM) around a person's obtained test score to represent the zone of the person's true score has been known since at least 1950.[13]
5. It is a fact that the use of SEM to produce a range (or zone of scores) to bracket an obtained IQ score is recognized and recommended as professional best practice in the diagnosis of MR/ID by the two professional associations cited in *Atkins v Virginia* as the sources to be used to define MR/ID.
6. It is a fact that the guiding psychometric constitution of standards for psychological and education testing (the Joint Test Standards) recognizes and recommends the use of the SEM in proper interpretation of test scores. It is a fact that these professional standards have been recognized vis-à-vis the collaboration of the American Psychological Association (APA), American Education Research Association (AERA), and National Council on Measurement in Education (NCME) since 1966 (AERA, APA, NCME, 1996).
7. It is a fact that all major academic and professional admissions examinations (GRE, SAT, ACT, LSAT, MCAT, DAT, state school achievement tests, etc.) use and report the tests SEM with scores. The application of the SEM measurement concept is not unique to IQ tests. The failure to accept the notion of SEM of IQ scores in Atkins cases is simply wrong and ignores the scientific facts and evidence from the well-developed field of psychometrics and the technology of standardized IQ testing. Those involved in Atkins MR/ID decisions, where IQ scores are one prong of the definition of MR, should not run from or misuse the concept of SEM. The psychometric concept of SEM should be embraced by the courts since the real value in understanding and appropriately interpreting the SEM is that it "provides a counteractive force to overinterpretation and overtrust of test data" (Feldt & Brennan, 1989, p. 106). ***SEM is a crime-blind or neutral statistic*** that should help the courts, when properly applied, to move away from the bright-line cut-off IQ score of 70 to a "*bright zone*" of allowable IQ scores when considering

[12] See table in front cover of Sattler (2001).

[13] Gulliksen (1950). *Theory of mental tests.* As discussed in Harvill (1991),

the diagnosis of MR/ID. This point has been echoed by one of the leading experts in intelligence test development and interpretation (Alan Kaufman) when he stated "what is vital is to internalize the reality that we don't earn a specific IQ—we actually earn a *range of IQ's* that most likely includes our true IQ" (Kaufman, 2009, p. 143).

8. Finally, the standard "error" should not be confused with the everyday use of the word "error" which suggests something avoidable (a fixable mistake).[14] The SEM is not avoidable, nor is it a fixable mistake. *The everyday* error that does need fixing occurs when the simple psychometric concept of IQ SEM-based confidence intervals or bands is ignored, abused, or misinterpreted. The SEM must be applied when evaluating the precision of IQ scores in Atkins MR/ID death penalty decisions. If the concept of fairness and equal protection holds true, the SEM of IQ scores must be applied to the scores of individuals under consideration for a diagnosis of MR/ID in Atkins cases. To not correctly apply this recognized and accepted measurement concept to this select population, while it prevails as standard practice for almost all reputable admissions and selection tests for other groups, is wrong. Period.

References

American Association on Intellectual and Developmental Disabilities (2010). *Intellectual Disability: Definition, classification, and systems of supports*. Washington, DC. Author

American Educational Research Association, American Psychological Association, and National Council on Measurement in Educations (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Kaufman, A. S. (2009). *IQ testing 101*. New York: Springer.

Law School Admission Council (1997), *What is a score band?* Author.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement—3rd edition*. New York: American Council on Education: Macmillan Publishing Company.

Green, B. F. (1981). In defense of measurement. *American Psychologist, 33*, 1001-1011.

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

Shavelson, R. J. (1996). *Statistical reasoning for the behavioral sciences—3rd Edition.* Boston: Allyn and Bacon.

[14] As described by Feldt and Brennan (1989), the term "error", in ever day conversation suggests an avoidable or correctable mistake. This is not the meaning of error in the context of standardized psychological testing.

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).