IAP AP 101

APPLIED PSYCHOMETRICS 101:

#11: Time to Stop Executing the Mentally Retarded – The Case for Applying the Standard Error of Measurement

In *Atkins v. Virginia*, the Supreme held that, under the Eighth Amendment, the death penalty is cruel and unusual punishment when applied to a mentally retarded individual. The diagnosis of mental retardation typically requires proof of three prongs or elements: significantly subaverage intellectual functioning, deficits in adaptive functioning, and onset of the condition in the developmental period. This article is concerned with the first prong of the diagnosis - significantly subaverage intellectual functioning, which usually means a measured IQ score that is two or more standard deviations below the mean. Since no IQ test is perfect, test developers determine and publish standard errors of measurement (SEM) for their tests, and leading clinical definitions of the mental retardation diagnosis encourage the use of the SEM to account for the inherent imprecision of the tests. However, despite the widespread acceptance of the SEM, some courts and legislatures have refused to consider the standard error of measurement (SEM) in *Atkins* cases, and some that do apply it in an irrational manner. As a consequence, some mentally retarded individuals still face the death penalty – despite the fact that the Supreme Court held that execution of the mentally retarded is unconstitutional.

Kevin F. Foley
Kevin S. McGrew

© Institute for Applied Psychometrics llc (IAP) 04-04-11 ver1

Time to Stop Executing the Mentally Retarded – The Case for Applying the Standard Error of Measurement

Kevin F. Foley[1] and Kevin S. McGrew[2]

Introduction. In *Atkins v. Virginia*, the Supreme held that, under the Eighth Amendment, the death penalty is cruel and unusual punishment when applied to a mentally retarded individual.[3] The diagnosis of mental retardation typically requires proof of three prongs or elements: significantly subaverage intellectual functioning, deficits in adaptive functioning, and onset of the condition in the developmental period. This article is concerned with the first prong of the diagnosis - significantly subaverage intellectual functioning, which usually means a measured IQ score that is two or more standard deviations below the mean. Since no IQ test is perfect, test developers determine and publish standard errors of measurement (SEM) for their tests, and leading clinical definitions of the mental retardation diagnosis encourage the use of the SEM to account for the inherent imprecision of the tests. However, despite the widespread acceptance of the SEM, some courts and legislatures have refused to consider the standard error of measurement (SEM) in *Atkins* cases, and some that do apply it in an irrational manner. As a consequence, some mentally retarded individuals still face the death penalty – despite the fact that the Supreme Court held that execution of the mentally retarded is unconstitutional.[4]

What Did Atkins Say, and What is Required to Comply With the Eighth Amendment? In discussing the concept of mental retardation, the *Atkins* Court looked to two clinical definitions – the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) and the definition provided by the American Association on Mental Retardation (AAMR, now known as the American Association on Intellectual and Developmental Disabilities [AAIDD]). The Court noted that, in general, the accepted clinical definitions of mental retardation have three requirements, (1) subaverage intellectual functioning, (2) deficits in adaptive functioning, and (3) onset before age 18. "[C]linical definitions of mental retardation require not only subaverage intellectual functioning, but also significant limitations in adaptive skills such as communication, self-care, and self-direction that became manifest before age 18."[5]

In what some might consider an unfortunate piece of law-making, the Court decided to leave it up to the states to define mental retardation and the procedures for implementing the constitutional ban.[6] "Not all people who claim to be mentally retarded

[1] Federal Administrative Law Judge. The views of Judge Foley as expressed in this paper are made in his individual capacity and should not be considered to be the policy or viewpoints of either the Social Security Administration or the U.S. Government.

[2] Dr. McGrew is the Director of the Institute for Applied Psychometrics (IAP).

[3] 536 U.S. 304 (2002).

[4] Although many professionals now refer to this population as intellectually disabled, this article will use the moniker used in the majority of the cases – mentally retarded.

[5] *Id*. at (slip copy, pg. 13).

[6] According to one commentator, "Like *Ford*, the Court in *Atkins* failed to articulate an adequate standard to ensure that the states will provide mentally retarded defendants the protection of the Eighth Amendment.

will be so impaired as to fall within the range of mentally retarded offenders about whom there is a national consensus. As was our approach in *Ford* v. *Wainwright*, with regard to insanity, 'we leave to the State[s] the task of developing appropriate ways to enforce the constitutional restriction'".[7] But insanity is different than mental retardation. "Insanity" for purposes of guilt or competence to be executed involves a legal concept, not a clinical diagnosis such as mental retardation. As the *Ford* Court noted, the concern was with "one whose mental illness prevents him from comprehending the reasons for the penalty or its implications",[8] not with whether the individual fits within any particular diagnostic pigeon-hole. "Mental illness" is broad; mental retardation is specific. There lies the rub.

Quite frankly, the *Atkins* Court erred when it blended together the procedural aspects of *Atkins* claims with the substance of a mental retardation diagnosis when it said that it was leaving to the States "the task of developing appropriate ways to enforce the constitutional restriction". It is one thing to allow states the choice of having its *Atkins* claims decided pre-trial rather than as a part of the sentencing proceeding, or having the judge decide the mental retardation issue instead of a jury. But it is another thing to allow states to monkey with the two major definitions of the diagnosis in an effort deny more cases than they should.

At the time that the Texas Court of Criminal Appeals decided the *Atkins* case of *Ex parte Briseno*,[9] the Texas legislature had not enacted a statute to cover mental retardation claims by persons facing the death penalty. Thus, the Texas high court for criminal matters was faced with developing such a standard on its own. If we could step into the shoes of the Texas Court of Criminal Appeals at the time it was deciding *Briseno*, what resources would we have to determine the standard to apply? We would

Variations among the states' definitions of mental retardation create extreme inconsistencies in the determination of whether an offender is recognized as mentally retarded." Cynthia A. Orpen, *Following in the Footsteps of Ford: Mental Retardation and Capital Punishment Post-Atkins*, 65 U. Pitt. L. Rev. 83, 91 (2003). Similarly, the court in *Maldonado v. Thaler*, 662 F. Supp. 2d 684 , No. H-07-2984 (S.D. Tex., Sept. 24, 2009) (slip copy at 36), *aff'd.,* 625 F. 3d 229 (5th Cir. 2010), *quoting Bell v. Cockrell*, 310 F. 3d 330, 332 (5th Cir. 2002), pointed out that, "A 'welter of uncertainty' followed the *Atkins* decision because '[t]he Supreme Court neither conclusively defined mental retardation nor provided guidance on how its ruling should be applied to prisoners already convicted of capital murder'", and another commentator claimed, "Although *Atkins* was supposed to help protect the rights of mentally retarded defendants . . . the manner in which states such as Tennessee have implemented *Atkins* has hindered them." Pramila A. Kamath, *Blinded by the Bright-Line: Problems With Strict Construction of the Criteria for Death Penalty Exemption on the Basis of Mental Retardation*, 77 U.Cin. L. Rev. 321, 343 (2008).

[7] *Atkins*, *quoting from*, *Ford v. Wainright*, 477 U.S. 399 (1986).

[8] *Ford*, 477 U.S. at 417. *Ford* was no model of clarity, however, and the legal standard set by the Court in that decision created its own set of problems. "The decision in *Ford* left open many questions about the legal standard." Jeffery L. Kirchmeier, *The Undiscovered Country: Execution Competency & Comprehending Death*, 89 Kentucky L.J. 263, 278 (2009/2010). First, it was Justice Powell's concurrence that was considered to contain the statement of the legal standard, not the plurality opinion. "Subsequent opinions from lower courts assumed that Justice Powell's definition of competence was controlling without much reference to the other language in *Ford*." Memorandum Opinion and Order, *Billiot v. Epps*, 671 F. Supp. 2d 840, Case No. 1:86CV549TSL, U.S. Dist. Ct. (S.D. Miss., Nov. 3, 2009), slip op. at 11. Second, and notwithstanding the Court's more recent competency decision of *Panetti v. Quarterman*, 551 U.S. 930 (2007), "courts today are still struggling with the reasons for the ban and how to define when one is incompetent to be executed." Kirchmeier, *supra* at 264.

[9] 135 S.W. 3d 1 (Tex. Cim. App. 2004).

start with the *Atkins* decision itself, which in turn would direct us to the DSM-IV and the AAMR standards. Too, we would be directed to the extant state statutes which define mental retardation and any federal standards and statutes as well. But this exercise can be a dangerous one, as explained later in the discussion of the Alabama case of *Smith v. State*.

The Indiana Supreme Court attempted to determine the parameters of *Atkins*. "We agree that *Atkins* did not provide for a uniform definition of mental retardation, but note that *Atkins* cited with approval the clinical definitions of mental retardation, explaining that while state statutory prohibitions against executing the mentally retarded are not uniform, all, including Indiana's, 'generally conform' to the clinical definitions."[10] The Indiana court added,

> "Although *Atkins* recognized the possibility of varying state standards of mental retardation, the grounding of the prohibition in the Federal Constitution implies that there must be at least a nationwide minimum. The Eighth Amendment must have the same content in all United States jurisdictions. Accordingly, we conclude that states are free to impose a higher standard, but the minimum definition of mental retardation sufficient to meet the national consensus found in *Atkins* must be uniform. Because *Atkins* explains that state statutes that provided the 'national consensus' against the execution of the mentally retarded 'generally conform' to the AAMR or DSM-IV definitions, we conclude that *Atkins* requires at least general conformity with those clinical definitions, but allows considerable latitude within that range. We agree with the concurrences that the Supreme Court specifically disavowed endorsement of the DSM-IV or any specific definition of mental retardation, but we think that the prohibition of the execution of the mentally retarded must have some content. There may be some flexibility in determining mental retardation, but we think that if a state's definition of mental retardation were completely at odds with definitions accepted by those with expertise in the field the definition would not satisfy the prohibition."[11]

The Indiana court failed to answer the most important questions – Can there be a bright-line cutoff, and if so, how low can a state place it before offending the Eighth Amendment? Must states apply the standard error of measurement (SEM) and the Flynn Effect?[12] Must an adaptive behavior instrument be used? What must be shown before age 18 (i.e., a diagnosis of mental retardation, an IQ score of 70 or below, or just some evidence consistent with mental retardation)? Must malingering be ruled out and, if so, how? Although the *Pruitt* court held that that the state could not require that the

[10] *Pruitt v. State*, 834 N.E. 2d 90, 107 (Ind. 2005), *cert. denied*, 548 U.S. 910 (2006).

[11] *Id*. at 108.

[12] The Flynn Effect refers to the phenomenon of rising IQ scores over time, estimated to be approximately 0.3 points per year. *See* J.R. Flynn, *Tethering the Elephant: Capital Cases, IQ and the Flynn Effect*, 12 Psychol., Pub. Pol. & Law 170 (May, 2006).

defendant prove mental retardation by the difficult “clear and convincing evidence” standard,[13] the Arizona Supreme Court reached just the opposite result.[14] This is an interesting conflict, considering that the Indiana high court reached its conclusion on the “clear and convincing evidence” issue using federal constitutional law.[15]

The Indiana court stated that there is “at least a nationwide minimum”; states could “impose a higher standard”; any standard “must have some content”; and whatever standard a state employs cannot be “completely at odds with definitions accepted by those with expertise in the field”. As implied by the Indiana Supreme Court, the courts and legislatures have no unique insight into the diagnosis of mental retardation, beyond what they learn from researchers, clinicians and the extant body of professional and scientific literature. Thus, the states’ definitions must conform closely to the accepted clinical definitions on all issues that could have a significant impact on who should be construed to be mentally retarded.[16] But how far can states go in determining what specific IQ score will qualify for a finding of mental retardation? When an attorney representing the state of Indiana before Seventh Circuit Court of appeals made the absurd argument that the defendant must show that he has an IQ of 60 or less, the court rejected such a contention out of hand. But the court would not say what the cutoff is between retarded and not retarded, only that it is “well above 60.”

> “At oral argument, the State argued that if we remanded Allen’s case for an *Atkins* hearing, the district court would need to determine whether Allen satisfies the ‘clinical definition’ of mental retardation (using Indiana’s standard) and *also* ‘whether or not Allen is among the class of offenders of which there is a total national consensus.’ This national consensus, according to the State, requires the district court to find that Allen has an IQ of 60 or below.
>
> “We reject this argument. Contrary to the State's assertion, the Supreme Court in *Atkins* did *not* establish a national standard for mental retardation but expressly left to the states the task of defining mental retardation. And to the extent that the Court acknowledged any limiting IQ score, that score was well above 60 (noting that an IQ score ‘between 70 and 75 or lower . . . is typically considered the cutoff IQ score for the intellectual functioning prong of the mental retardation definition.’).”[17]

Arizona has an unusual statutory scheme, with different burdens depending on IQ level.

[13] *Id*. at 103.

[14] *State v. Grell*, 212 Ariz. 516, 525, 135 P. 3d 696 (Ariz. 2006), *cert denied*, 550 U.S. 937 (2007).

[15] *Pruitt*, *supra* note 10 at 103.

[16] In *Pruitt*, the Indiana Supreme Court approved the state’s adaptive behavior prong of the statutory definition of mental retardation, only because the court found that “it is very similar to the revised AAMR definition, and therefore within the range of permissible standards under the Eighth Amendment.” *Id*. at 108.

[17] *Allen v. Buss*, 558 F. 3d 657, 665 (7th Cir. 2009) (citations omitted).

> "State procedures must ensure that those about whom there is national consensus are protected from execution, but left states otherwise free to craft their laws for determining which defendants meet the consensus standard. By providing differing procedures based on the defendant's IQ, Arizona law reflects this concept. Those with IQ scores of 65 or below face a comparatively lower bar, while those whose IQ scores suggest greater intelligence must go to greater lengths to prove their mental retardation. The legislature placed a heavier burden on those who do not fall within the group about whom there is national consensus regarding their right not to be executed."[18]

But again, "the group about whom there is national consensus" was not specified by the court; thus the parameters of this group remain elusive. The only real consensus that can be discerned from the states' treatment of the *Atkins* claims is that the definitions provided by the DSM-IV and the AAMR (now the AAIDD) are the crucial focus and these definitions should control the goings-on in *Atkins* cases more than any other definition. In general, the states borrowed generously from the AAMR in describing the intellectual functioning prong of their statutory definitions.[19] State supreme court decisions only further demonstrate the importance and relevance of the prevailing clinical definitions. In discussing its state's standard, the Mississippi Supreme Court stated that, "These definitions [the DSM and AAMR] were previously adopted and approved by this Court".[20] The Pennsylvania Supreme Court, "held that a defendant may establish mental retardation through resort to either the AAMR's Mental Retardation standard or the American Psychiatric Association standard set forth in the Diagnostic and Statistical Manual of Mental Disorders (4th ed. 1992) ('DSM-IV')."[21] According to the Texas Court of Criminal Appeals, the highest court in that state for criminal matters, "We have adopted the American Association on Mental Retardation (AAMR) definition of mental retardation for *Atkins* claims presented in Texas death-penalty cases."[22] In rejecting a challenge to Louisiana's mental retardation "onset before age 18"

[18] *State v. Grell*, 212 Ariz. 516, 525, 135 P. 3d 596 (Ariz. 2006).

[19] The Death Penalty Information Center provides a summary of state statutes in effect at the time that *Atkins* was decided, as well as laws enacted to comply with *Atkins*. At the time *Atkins* was decided, for the intellectual functioning prong, most states with such statutes required evidence of "significantly subaverage intellectual functioning" or "significantly subaverage general intellectual functioning". Some states – Arizona, Colorado, Connecticut, Florida, Indiana, Kansas, Missouri and New York - did not define these phrases further. Some states – Maryland, Kentucky, North Carolina, Tennessee and Washington - specified an IQ of 70 or less, and other states' statutes – South Dakota, Nebraska, New Mexico and Arkansas – spoke in terms of a number (i.e., 70) providing a ***presumption*** of mental retardation. *State Statutes Prohibiting the Death Penalty for People with Mental Retardation*, Death Penalty Information Center, http://www.deathpenaltyinfo.org/state-statutes-prohibiting-death-penalty-people-mental-retardation (accessed Nov. 22, 2009).

[20] *King v. State*, 2007-DR-01336-SCT (Miss., Sept. 24, 2009), 2009 Miss. LEXIS 460, *reh. denied*, 2010 Miss. LEXIS 4.

[21] *Commonwealth v. Vandivner*, 599 Pa. 617, 644, 962 A. 2d 1170 (2009), *cert. denied*, 130 S. Ct. 260 (2010).

[22] *Williams v. State*, 270 S.W. 3d 112, 113 (Tex. Crim. App. 2008).

requirement, the Louisiana Supreme Court pointed to how "the provision that the onset of mental retardation manifest by age 18 comports with *Atkins*, the American Association of Mental Retardation (AAMR), *see* AAMR, Mental Retardation, p. 1 (10th ed. 2002), the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM)".[23] The Kentucky Supreme Court noted with approval its state's definition, which is "defined by the same three criteria established by the AAMR and the American Psychiatric Association and approved in *Atkins*".[24] And, of course, it was the AAMR and the DSM-IV that the Supreme Court looked to in *Atkins*.

Texas – Hedging Its Bets. As mentioned above, in *Williams*, the Texas high court claimed that it had adopted the AAMR definition for resolving *Atkins* claims. But, a recent federal district court case from Texas makes the point that Texas has really not adopted the AAMR definition. Putting aside the Texas-court-created adaptive behavior criteria fiasco for the moment, Texas apparently has not resolved the IQ level issue, meaning it has not adopted the AAMR criteria for determining an individual's level of intellectual functioning. The AAMR (now the AAIDD) has taken the position for some time that the SEM should be taken into consideration when making the diagnosis of mental retardation. Since the Texas legislature has never weighed in on the issue, the Texas Court of Criminal Appeals had to create a legal standard for resolving *Atkins* claims in Texas state courts. It did so in *Ex parte Briseno*, where the court stated,

> "Most Texas citizens might agree that Steinbeck's Lennie should, by virtue of his lack of reasoning ability and adaptive skills, be exempt [from execution]. But, does a consensus of Texas citizens agree that all persons who might legitimately qualify for assistance under the social services definition of mental retardation be exempt from an otherwise constitutional penalty? Put another way, is there a national or Texas consensus that all of those persons whom the mental health profession might diagnose as meeting the criteria for mental retardation are automatically less morally culpable than those who just barely miss meeting those criteria? Is there, and should there be, a 'mental retardation' bright-line exemption from our state's maximum statutory punishment?"[25]

The federal district court in *Maldonado v. Thaler* stated that, "While Texas has not established 70 as a bright-line standard, it has not expressly adopted another score as the retardation threshold either."[26] Apparently, the Texas high court for criminal matters feels it needs some guidance from others in order to do so. "[T]he Court of Criminal Appeals has refused to 'answer that normative question without significantly greater assistance from the [Texas] citizenry acting through its Legislature.'"[27]

[23] *State v. Anderson*, 996 So. 2d 973, 986 (La. 2008), *cert. denied*, *Anderson v. Louisiana*, 129 S. Ct. 1906 (2009). The Tennessee Supreme Court has held similarly. *State v. Strode*, 232 S.W. 3d 1 (2007).

[24] *Bowling v. Commonwealth*, 163 S.W. 3d 361, 370 (Ky. 2005).

[25] *Ex parte Briseno*, 135 S.W. 3d 1, 6 (Tex Crim. App. 2004).

[26] *Maldonado v. Thaler*, *supra* note 6, slip op. at 40.

[27] *Id*. (slip copy at 40), *quoting Ex parte Briseno*, 135 S.W. 3d at 6.

Quite frankly, the Texas "standard" is a cop-out. If the legislature neglects to enact an appropriate statutory standard, then the court *must* do so. Texas litigants are entitled to have their day in court, and their claims promptly resolved. The Texas Court of Criminal Appeals cannot claim to adopt a standard like the AAMR definition, then refuse to apply that standard, or apply only the parts of it that result in more claims being denied. The only reason for failing to fully adopt the AAMR standard is to deny more *Atkins* claims. In fact, Texas' failure to come to grips with the SEM has caused Texas to be one of the states that has paid lip service to the concept, with consequent irrational results.

Understanding the Standard Error of Measurement and Its Importance in Achieving a Fair Result in *Atkins* Cases.

Introduction to Measurement Error and the SEM. Four times a year the future lawyers and judges of our country anxiously await the receipt of their LSAT (Law School Admission Test)[28] examination results. Given the prominent role LSAT scaled scores (scores range from 120-180, average of approximately 151) [29] 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 it is 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:[30]

> "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."

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

[29] Average LSAT Scores for Top Law Schools, http://www.lsatprepcourse.com/law.htm (accessed Dec. 13, 2009)

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

For a young person who dreams of attending Harvard Law School, a score of approximately 170 or above is desirable.[31] 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 (rounded to 3 for ease of presentation).[32] As will be explained later, the incorporation of the LSAT SEM-based confidence band allows the 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 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 admissions 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 [33] 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), it should follow that the IQ scores for individuals being considered for a diagnosis of mental retardation (and potential capital punishment) be accorded the same protections provided to others under the professionally and scientifically grounded constitution of psychological measurement (i.e., the Joint Test Standards).[34]

Unfortunately, a portion of the misunderstanding and misapplication of the IQ SEM rule in mental retardation diagnosis may be due to the construct and statistically dense measurement and statistical concepts often used by psychometric specialists to explain reliability and SEM, which include the following: Hypothetical true score. Observed or measured score. Reliability. Different types of reliabilities. Standard deviation. Standard error. SEM. Conditional SEM. SEM-based confidence interval

[31] The average (between 25th and 75th percentile ranks) Harvard LSAT score is 170-175. Average LSAT Scores for Top Law Schools, *supra* note___.

[32] Law School Admission Council (1997), *What is a score band?*

[33] For illustrative purposes we 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.

[34] 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.

around the observed score versus SEM-based confidence interval around the true score. Symmetrical versus asymmetrical confidence bands. 68 %, 90%, 95% confidence intervals. Ability-centered, local, conditional, or overall reliability and SEM. $\mathrm{SEM} = \mathrm{SD} * \sqrt{1 - \mathrm{r}(11)}$.

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.[35] One way to resolve the complexity problem is to explain the 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 using 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 an 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 just 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 generally 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"[36] and the important statistical estimate derived from a test's reliability—the SEM (also called "standard error").

[35] The reader interested in the various nuances of reliability theory and the standard error of measurement are urged to read two relatively concise instructional modules produced by the National Council on Measurement in Education (NCME). The module *Understanding Reliability*, Ross E. Traub & Glenn E. Rowley, authors (1991), is available for download at http://www.ncme.org/pubs/items/15.pdf (accessed Dec. 13, 2009). The *Standard error of measurement* module, Leo M. Harvill, author, (1991) is available for download at http://www.ncme.org/pubs/items/16.pdf (accessed Dec. 13, 2009). The complete set of NCME instructional modules that address an array of psychometric topics can be found at http://www.ncme.org/pubs/items.cfm. Although there are some technical differences in how reliabilities and the 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 their SEM.

[36] Leonard Feldt & Robert Brennan , *Reliability*, in Educational measurement—3rd edition 105 (R. L. Linn ed., 1989), Also, the material immediately preceding this quote is based on ideas and examples from Feldt & Brennan.

The quantification and objective comparison of different individuals in human performance situations is made possible through the use of standard sets of tasks and measurement procedures. For example, when measuring bowling performance, each "test" session is the same - constant 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, which allows each person's bowling ability to be judged using 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."[37] 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 using 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.[38] 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."[39] 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 through a common method or mental yardstick.[40]

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 (SEM) has been developed to quantify the amount of potential error inherent 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

[37] Bert F. Green, *In defense of measurement*, 33 Am. Psychol. 1001 (1978).

[38] Sattler (2001), Assessment of Children: Cognitive Applications—4th Edition, at p. 4.

[39] American Educational Research Association, American Psychological Association, and National Council on Measurement in Educations, Standards for educational and psychological testing 182 (1999).

[40] Green, supra note 37.

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 “glump” 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 “glump.” 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 experts’ imaginations.[41] It is never possible (nor would it be humane or ethical) to administer the same IQ test 100 times to one individual in a relatively short span of time. 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).

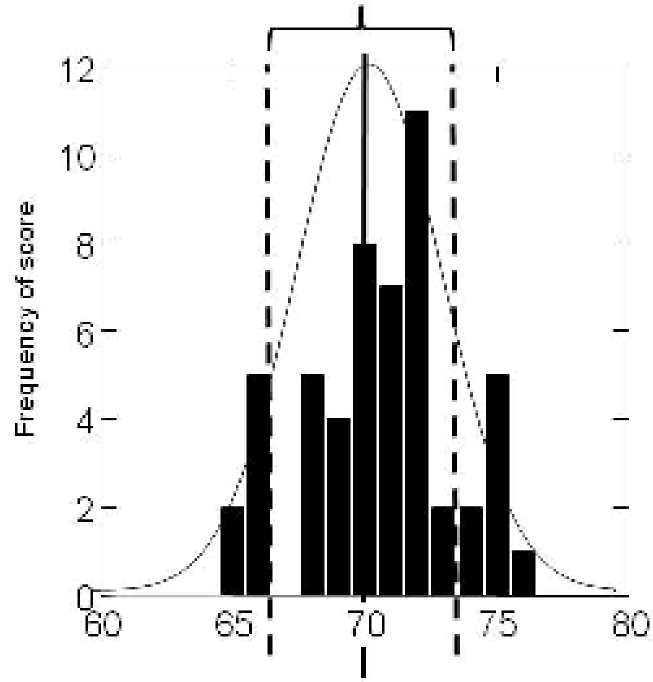
A Real-World Example of SEM. Rather than engaging in quantoid-speak and the temptation to demonstrate an expertise in statistical symbol manipulation and equation wizardry (and bore the reader to sleep), we instead provide a real-world example of the measurement concepts of reliability and SEM. The example 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) are presented in Figure 1.

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,

[41] Alan S. Kaufman, IQ testing 101 at 141 (2009).

during their professional education, took one or more courses in statistics—the normal distribution or bell curve.[42] 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 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 "glumping" 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.[43] 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

Standard deviation (approx 3 points) of all 52 scores is the person's SEM (standard error of measurement). We can be 68 % confident (confidence band of ±3 pts) that this persons true ability score is between 67 and 73 (vertical dashed lines).



Solid vertical line is average of the 52 scores and is the best estimate of the person's true ability score

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

[42] 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.

[43] Shavelson (1996), Statistical reasoning for the behavioral sciences—3rd Edition.

they operate randomly in either direction, will form a distribution that approximates the normal curve (as illustrated in Figure 1). [44]

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). This score is 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 (we round 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 SEM (SEM).[45] 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.[46] This range of scores is known in statistical terms as the 68 % confidence band or confidence interval for the ability score.[47] Using the known mathematical characteristics of the normal curve, we can be 95% confident that the person's true ability score will fall from -

[44] Kaufman (2009), IQ testing 101 at p.141.

[45] Richard J. Shavelson, Statistical reasoning for the behavioral sciences (3rd ed. 1996).

[46] Throughout this paper we use the phrase (or some variation) that the standard error of measurement allows us to be "__% confident that the persons true ability score would be somewhere in the range of __ to __. One of the standard error of measurement technical nuances referenced by Harvill, supra note 35, is whether confidence bands should be constructed around (a) a calculated true IQ score to estimate what an obtained or measured score would be, or (b) an obtained or measured score to estimate what a true score would be. Slightly different statistical formulas and procedures are used to calculate the standard error of measurement depending on whether approach "a" or "b" is used. As noted by Harvill, supra note 35, these subtle nuances are often misunderstood and misstated in measurement books. This nuance reflects one of those issues previously described as "slicing of the measurement and statistical butter with a laser beam" among psychometricians that has minimal impact on the most common professional standard error of measurement confidence band methods employed by psychologists. The most commonly used procedure (adding standard error of measurement to and subtracting it from the obtained score to estimate the true score", Harvill, supra note 35, is generally a satisfactory estimate within "reasonable limits." We could qualify all our confidence band statements in this article with language such as the confidence band reflecting "reasonable limits" or the confidence band provides a "good approximation." We choose not to add these nuanced qualifications in order to reduce potential confusion in comprehension of the text and to maintain language that is most commonly used when discussing standard error of measurement-based confidence bands.

[47] Jerome M. Sattler (2001), Assessment of Children: Cognitive Applications 109 (4th ed. 2001).

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, it can be determined from the test's calculated reliability, the average SEM across individuals.[48] 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 a professionally recommended best practice (when making mental retardation 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 mental retardation classification literature,[49] and almost always when discussed in an *Atkins* mental retardation determination, the rule of thumb of ± 5 IQ points (for a confidence band that spans 10 IQ points) is recommended.

Before proceeding, we must confess to some necessary deliberate deceit, 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 that golfer Tiger Woods achieved on the same golf course (Augusta National Golf Club—home of the annual Masters golf tournament). Granting the authors 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[50] 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 Woods did not

[48] Traub & Rowley, *supra* note 35.

[49] American Association on Intellectual and Developmental Disabilities (2010), *Intellectual disability: Definition, classification, and systems of supports—11th Edition* at p. 36.

[50] 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.

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 of scores in the middle of the histogram (the "glumping" of scores) and the more extreme high and low scores occurring less frequently. We can state that the best estimate of Tiger Woods' true Masters golfing ability is 70. More importantly, we recognize 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 Wood's scores across repeated "testing." And, we can quantify his personal Masters SEM (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. 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.[51] 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.

Summary of SEM for Legal Fact Finders. In this concluding section on SEM, we offer the following facts 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." [52] Our example of 52 rounds of golf by Tiger Woods on the Masters course (test) is a small sample demonstration of this concept.

(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.

[51] These factors represent a synthesis of factors mentioned by Feldt & Brennan, *supra* note 36 at 107 and Sattler, *supra* note 38 at 105.

[52] Kaufman, supra note 41 at 141.

(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.[53] It is also a fact that the mathematical assumptions, logic and methods 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. [54]

(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 mental retardation by the two professional associations cited in *Atkins v Virginia* as the sources the Supreme Court looked to in order to define al retardation.

(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.[55]

(7) It is also 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* mental retardation decisions, where IQ scores are one prong of the definition of mental retardation, 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."[56] SEM is a crime-blind and neutral statistic that should help the courts, when properly applied, to move

[53] Sattler, supra note___, from table inside of front book cover (Historical Landmarks in Cognitive and Educational Achievement).

[54] H. Gulliksen Theory of mental tests (1950), as discussed in Harvill, supra note 35.

[55] American Educational Research Association, American Psychological Association, and National Council on Measurement in Educations , Standards for educational and psychological testing, 182 (1999).

[56] Feldt & Brennan (1989). Reliability. In R. L.. Linn (Ed.), *Educational measurement—3rd edition*, at p. 106.

away from the bright-line cut-off IQ score of 70 to a "bright zone" of allowable IQ scores when considering the diagnosis of mental retardation. 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." [57]

(8) Finally, the standard "error" should not be confused with the everyday use of the word "error" which suggests something avoidable (a fixable mistake).[58] 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* mental retardation death penalty decisions. 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.

Paying Lip Service to the Standard Error of Measurement with Irrational Results. Some state supreme courts that have held that there must be an IQ score of 70 or less, and that the standard error of measurement will not be applied.[59] While this result is unfortunate and inconsistent with the prevailing treatment of IQ test results by psychologists, at least one can understand such misplaced holdings, especially where there is a state statute that specifies a score of 70 or less. But in such states, there is a concern that some mentally retarded individuals will be executed, despite the constitutional ban. After discussing its state's cutoff score of 70, an appeals court in Tennessee stated, "Unfortunately, by refusing to consider ranges of error, it is our view that some mentally retarded defendants are likely to be executed in Tennessee, particularly in a case similar to this one where the defendant's I.Q. is so close to the brigh-line (sic) cutoff of 70."[60]

What is not understandable or acceptable is the spin some courts use in an effort to reject application of the standard error of measurement in *Atkins* cases. In the recent Texas case involving Bobby Wayne Woods, the Texas Court of Criminal Appeals offered this explanation for concluding that Woods did not have IQ scores sufficient for a finding of mental retardation:

> "[T]aking into account [psychologist] Schmitt's testimony concerning a 'possible 5 point deviation on either side,' applicant's

[57] Kaufman (2009), *IQ testing 101* at p.143.

[58] As described by Feldt and Brennan (1989) at p 105, 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.

[59] *See, e.g., Cherry v. State*, 959 So. 2d 702 (Fla.), *cert. denied*, 552 U.S. 993 (2007) and *Pizzuto v. State*, 202 P. 3d 642 (Idaho 2008).

[60] *Cribbs v. State*, Tenn. Crim Ct. App., Case No. W2006-01381-CCA-R3-PD (July 1, 2009) (slip op., pgs.47-48), 2009 Tenn. Crim. App. LEXIS 524, *app. denied*, 2009 Tenn. LEXIS 881. Cribbs had full scale IQ scores of 70, 75, and 73, although one psychologist reported that he believed Cribbs' I.Q. was likely higher than the assessed score of 73.

> Full Scale IQ scores range anywhere from 63 to 78. [The psychologist's] adjusted IQ scores for applicant, therefore, establish that applicant's IQ may fall below 70. . . Even under [this] analysis, a rational finder of fact could find that applicant's Full Scale IQ falls above 70."[61]

A similar tack was taken by an Ohio appellate court in *State v. Lawson*,[62] where the court stated, "Ultimately, the [trial] court concluded that, even considering appellant's most recent IQ score only, appellant's 'true' score could be as low as 62.65-63.26 or as high as 72.65-73.26, when considering the standard error of measurement, and that the record showed an equal likelihood of his IQ being above 70 as below 70. Accordingly, the court concluded that appellant failed to prove the first criterion of mental retardation". The trial court was affirmed on appeal, meaning this approach was approved by the appeals court.[63] Similarly, the court in *Byrd v. State*,[64] noted, "By relying on the mere possibility that his true IQ falls at the low end of the confidence interval or, as he described it, the 'margin of error,' Byrd has not met his burden to establish that it is more likely than not that his IQ is 70 or below."[65]

The holdings of these cases are misplaced. Under the reasoning of these courts, an IQ score of 70 – or even 66 – is not enough to carry the day. That is because the "fact-finder" can simply choose, apparently without any explanation or reasoning, to apply the 5-point SEM to increase the obtained score to arrive at an "adjusted" IQ score above the cutoff of 70, even when the obtained score is a 66 or 67. Or, put another way, with a full scale score of, say, 70 on the WAIS-III, we have a 95% confidence interval that the true score falls somewhere between 65 and 75. But who is to say whether the true score is 65-70, or 71-75? It is just as likely to be within 71-75 as it is to be within 65-70. Thus, since the defendant has the burden of proof – usually by a preponderance of the evidence – he cannot satisfy his burden because it is not more probable that his score is within the 65-70 range.

That reasoning is pure spin. The concept of the SEM as applied to mental retardation assessment is intended to allow for a person to obtain the diagnosis of "mentally retarded" in the situation where he has the requisite deficits in adaptive functioning and onset before age 18, and an IQ score within 5 points of the cutoff of 70 points. As stated in the AAMR definition manual, ninth edition, "When the existence of intellectual limitations is uncertain or equivocal, the decision should be made that would

[61] *Ex Parte Woods*, AP-76,034 (Tex.Cr.App., Oct. 7, 2009), 2009 Tex. Crim. App. LEXIS 1432, *cert. denied*, 130 S. Ct. 794 (2009).

[62] 2008 Ohio 6066 (Ohio App.) *appeal denied*, 918 N.E. 2d 525 (Ohio 2009).

[63] However, a different Ohio appeals court stated, "a trial court cannot simply reject measurement error as a concept. In determining whether an individual is intellectually deficient, the AAMR and DSM-IV-TR include an assessment of measurement error in the test itself". *State v. Burke*, 2006 Ohio 1026 (Ohio App., unpub.), *rev. denied*, 109 Ohio St. 3d 1506 (Ohio 2006).

[64] Case No. CR 07-0113 (Ala. App. May 1, 2009).

[65] The court in *Ledford v. Head*, 2008 U.S. Dist. LEXIS 21635 (N.D. Ga.) used the same reasoning as the above cases when it concluded that there was "no basis for assuming a the standard error of measurement lowered [Ledford's] score enough to meet Georgia's mental retardation standard."

result in services that would be most advantageous to the individual."[66] Or, in other words, the individual gets the benefit of the doubt. The 5-point standard error of measurement was never intended to screw a defendant out of a diagnosis that he qualifies for under accepted assessment practices, except perhaps in some of these "tough on crime" states. But arriving at a fair and scientifically sound result in the *Atkins* cases is not about being easy or tough on crime. It is about making the best decision within the limitations of the available science and commensurate with the Eighth Amendment.

The Florida Supreme Court's decision in Joe Elton Nixon's case was nothing short intellectually dishonest. In the trial court, Nixon's attorneys presented evidence that Nixon had an IQ of 73, and that when one considered the Flynn Effect and the standard error of measurement, Nixon satisfied the statutory requirement of an IQ score that was two standard deviations below the mean. The trial judge ruled against Nixon on the basis that the Florida Supreme Court had set a bright line cut-off of 70 in the case of *Cherry v. State*.[67] According to the Florida Supreme Court's *Nixon* decision, "Nixon claims that in *Cherry*, we interpreted section 921.137(1) to create fact-finding procedures that preclude a defendant from presenting relevant material. Nothing in *Cherry* or section 921.137 precludes a defendant from presenting any evidence that is germane to the issues involved in a mental retardation claim."[68]

Nixon argued that by requiring an IQ score of 70 or below, and refusing to consider and apply the standard error of measurement and the Flynn Effect, "The Circuit Court thus applied Cherry to hold legally irrelevant evidence that both experts, and the Court itself, agreed was scientifically relevant. . . This legal preclusion from giving the scientific consensus evidentiary weight is unconstitutional".[69] The problem with the Florida Supreme Court's decision is, it claims that any defendant is free to present "any evidence that is germane to the issues involved in a mental retardation claim", which would ostensibly include evidence concerning the Flynn Effect and the standard error of measurement. But since the Florida Supreme Court has held that a score of 70 or below – period – is necessary, the court's language was mere surplusage. Specifically, the Florida Supreme Court stated in *Cherry* that, "The fundamental question considered by the circuit court and raised in this appeal is whether the rule and statute provide a strict cutoff of an IQ score of 70 in order to establish significantly subaverage intellectual functioning."[70] The court answered the question in the affirmative and held that Cherry's IQ score of 72 was inadequate. So, in Florida, you can present any evidence that is germane, including evidence of the standard error of measurement, but don't hold your breath, since – as argued by Nixon – such evidence is legally irrelevant. The Florida Supreme Court might as well have said, "Don't bother submitting such evidence, because Florida courts will give it no weight and no consideration whatsoever. A 70 is a 70 is a 70."

[66] AAMR Definition Manual (9th ed.) at 14.

[67] 959 So. 2d 702 (Fla.), *cert. denied*, 552 U.S. 993 (2007). Nixon argued that the Florida Supreme Court's *Cherry* decision violated both the Florida and United States constitutions. Reply Brief of Appellant, Nixon v. State, Case No. SC07-953 n.1, filed Feb. 9. 2008).

[68] *Nixon v. State*, 2 So. 3d 137 (Fla. 2009).

[69] Apellant's Reply Brief, *supra* note 67 at 9-10.

[70] 959 So. 2d at 712.

Looking to an unrelated statutory definition for guidance can be dangerous and the conclusions inappropriate if one is not familiar with the specifics of that area of the law. In refusing to apply a 5 point SEM an Alabama appellate court relied on a federal court case for the proposition that Social Security law does not apply the SEM, meaning without an IQ score of 70 or less the person is out of luck[71] However, the Alabama court was either ignorant of Social Security law or purposely presented only a partial picture of what happens when a Social Security adjudicator decides a case involving possible mild mental retardation. There is more than one way for a disability claimant to obtain relief on the basis of low cognitive functioning. One, he can show his condition *meets* a listing. In this case, he would need to show an IQ score of 70 or below.[72] However, unlike the AAMR and DSM-IV standard – which typically looks at the full-scale IQ score, the Social Security regulations permit the use of any of the three IQ scores (full-scale, performance and verbal) in determining whether the claimant meets the listing. Second, the Alabama court neglected to address federal case law in which at least one court held that the SEM should be applied in a Social Security disability case.[73] Third, there is a strong argument that, even if an IQ score of, say, 72 does not meet the listing, in consideration of the SEM, a score of 72 *equals* the listing, with the result that the person would be deemed to be disabled, assuming the other elements of the listing were proven. Social Security disability can be established by showing that the claimant *meets* or *equals* one of the listed impairments. Fourth, even if the claimant does not meet or equal the listing, he can still prevail if the evidence shows that in light of his impairments he cannot perform any of his prior jobs or any other jobs that exist in significant numbers.[74] Unlike the four ways of prevailing in a Social Security case, in a state that refuses to apply the SEM, the death penalty defendant loses, even if his obtained full scale IQ score is a 71. The Alabama court's inadequate statement of Social Security law was inexcusable.

Some of the courts' chicanery has not gone unchecked. In light of the *Atkins* Court's statement that it was leaving it up to the states to determine issues such as the application of the SEM, some courts have observed that, "'It would be wholly inappropriate for [a federal court], by judicial fiat, to tell the States how to conduct an inquiry into a defendant's mental retardation.'"[75] Yet, some courts are doing just that.

[71] *Smith v. State*, Ala. Ct. Crim App. No. 05-0561, Sept. 26, 2008, 2008 Ala. Crim. App. LEXIS 172.

[72] The mental retardation listing in the Social Security regulations, Listing 12.05 ,*et seq*, contains a specific subsection for mild mental retardation, 12.05 C. This subsection requires that the person meet the regulatory definition for mental retardation (IQ of 70 or less, deficits of adaptive functioning, and onset before age 22), plus have another impairment imposing work-related limitations in functioning. *See* 20 CFR § 404, Subpart P, appen. 1.

[73] *Beyerink v. Astrue*, N.D. Iowa, No. C07-3018-PAZ, Feb. 14, 2008, 2008 U.S. Dist. LEXIS 11098 ("Taking into account the five-point margin of error presumed by the testing protocol, Beyerink's Verbal IQ clearly falls within the range required by the Listing, despite the variation in the two test results.") While this court and the decision it cited to may have been incorrect in its application of federal disability law, the point is the Alabama court should have acknowledged the contrary authority and discussed why it should not apply.

[74] *See* 20 C.F.R. § 404.1520(a)(4), describing the SSA's 5-part sequential evaluation process.

[75] *Maldonado*, *supra* note 6 at slip op., pg. 36, *quoting from*, *In re Johnson*, 334 F.3d 403, 405 (5th Cir. 2003).

The court in *Thomas v. Allen*,[76] observed that the law in Alabama required an IQ score of 70 or below, notwithstanding that the five point standard error of measurement (SEM) and the so-called Flynn Effect (which some say allows for an adjustment of IQ scores of 0.3 points for every year since the test's norms were published) are commonly accepted scientific principles.[77] This did not keep the federal court in *Thomas* from re-writing existing state law.

> "[E]ven though the legal cut-off score for a finding of 'significantly subaverage intellectual functioning' is stated in opinions of the Alabama Supreme Court as 'an IQ of 70 or below,' a court should not look at a raw IQ score as a *precise* measurement of intellectual functioning. A court must also consider the Flynn effect and the standard error of measurement in determining whether a petitioner's IQ score falls within a *range* containing scores that are less than 70."[78]

In the federal appeals court case of *Moore v. Quarterman*,[79] the dissent lambasted the majority for allowing the district court to re-write Texas law: "Texas caselaw is explicit, though, in holding that an actual IQ between 70 and 75 is not sufficient for a finding of retardation." This is so, "*regardless* of the margin of error."[80]

The Fourth Circuit Court of Appeals in *Walker v. True*,[81] remanded the case to the district court to consider the application of both the Flynn Effect and the SEM, despite the fact that the Virginia Supreme Court has held that a score of 70 or below is necessary.[82]

[76] 2009 U.S. Dist. LEXIS 50825 (N.D. Ala.), *aff'd.*, 607 F.3d 749 (11th Cir. 2010).

[77] In Kentucky there is a "bright line" cutoff score of 70 or below. *Woodall v. Simpson*, 2009 U.S. Dist. LEXIS14328 (W.D. Ky.).

[78] *Id.*

[79] Dissenting Opinion, Case. No. 05-70038 (5th Cir., Aug.21, 2009) (unpub.), 2009 U.S. App. LEXIS 19015.

[80] *Id.* (emphasis in original).

[81] 399 F. 3d 315 (4th Cir. 2005).

[82] In *Johnson v. Commonwealth*, 591 S.E. 2d 47 (Va. 2004), *judgment vacated*, 544 U.S. 901 (2005), the court held that Johnson's IQ scores of 75 and 78 exceeded the statutory threshold of 70. A later panel of the federal appeals court stated that "neither *Atkins* nor Virginia law appears to require expressly that these theories be accounted for in determining mental retardation status", and the court tried to distinguish *Walker* on the basis that *Walker* was a *de novo* consideration of Walker's *Atkins* claim. But this assertion really doesn't make sense. The federal habeas court is required to apply the state's law, and if Virginia requires an IQ of 70 or below, period, then that is the law. It makes no sense to require a court to consider a scientific concept if the concept has been deemed to be irrelevant to the outcome. The later decision of the Fourth Circuit, *Green v. Johnson*, 515 F. 3d 290 (4th Cir.), *cert. denied*, 128 S. Ct. 2527 (2008), confuses the issue because, in this case, the appeals court held that the district court should have deferred to the Virginia Supreme Court's finding that "three of Green's four I.Q. test scores exceed the maximum score of 70". Two of Green's IQ scores were WAIS scores of 74 and the third score was an 84 on the Ammons & Ammons Quick Test. The so-called "Quick Test" should have been given little weight because it is a far cry from a full form IQ test, so the other scores should have provided a result similar to *Walker*. Johnson's death sentence was vacated by the U.S. Supreme Court and remanded back to the Virginia courts for consideration of the matter in light of *Roper v. Simmons*, 543 U.S. 551 (2005). *Johnson v. Virginia*, 544 U.S. 901 (2005). Insofar as Walker's case is concerned, on remand, the district court concluded that

So there we have it – We have a court like the Tennessee Court of Appeals lamenting that its state's Supreme Court's holding that a 70 or less is required, notwithstanding the SEM's widespread acceptance, will cause the untenable result that "some mentally retarded defendants are likely to be executed in Tennessee".[83] And we have some federal courts ignoring *Atkins'* instruction that it is the states' prerogative to set the parameters and procedures involving mental retardation claims by those facing death row. Surely, this is not what the U.S. Supreme Court envisioned when it issued its ban on the execution of the mentally retarded.

Conclusion. "In the course of recognizing the right in the Eighth Amendment of mentally retarded defendants not to be executed, the Supreme Court has identified that right as grounded in a fundamental principle of justice."[84] States must acknowledge and apply the SEM. Failure to do so can result in a 50 percent reduction in the number of persons who can qualify for a diagnosis or finding of mental retardation. Obviously, a concept which can have such a dramatic effect on the outcome of *Atkins* cases, and whose application can be based on nothing more than a court or legislature's desire to appear tough on crime, must be taken into consideration in all mental retardation claims by those facing the death penalty.

Walker was not mentally retarded, and on further appeal, the majority decision of the Fourth Circuit Court of Appeal did not even address IQ scores. "Although Walker devotes much of his appeal to the district court's analysis of the I.Q. prong, it is unnecessary for us to address those arguments because we conclude that the court did not clearly err in rejecting his claim on the adaptive prong." *Walker v. Kelly*, 593 F. 3d 319, Case No. 06-23 (4th Cir., Jan 27, 2010), *cert. denied*, 176 L. Ed. 2d 1215, slip op. at 7. According to the concurring and dissenting opinion, however, on remand the district court felt that two 2006 decisions from the Fourth Circuit stood "for the proposition that courts should 'refus[e] to use the standard error of measurement to lower IQ scores in *Atkins* cases due to the inherent speculation of using the standard error of measurement to lower an IQ score when it could just as likely be used to raise an IQ score.'" *Id.*, slip op. at 32 (concurring and dissenting), quoting district court order.

[83] *Cribbs v. State*, Tenn. Crim Ct. App., Case No. W2006-01381-CCA-R3-PD (July 1, 2009) (slip op., pgs.47-48), 2009 Tenn. Crim. App. LEXIS 524, *appeal den.*, 2009 Tenn. LEXIS 881.

[84] *Pruitt*, *supra* note 10 at 101.