Salekin, K.L., Neal, T.M.S., & Hedge, K.A. (2018). Validity, inter-rater reliability, and measures of adaptive behavior: Concerns regarding the probative versus prejudicial value. Psychology, Public Policy, and Law, 24, 24-35. doi: 10.1037/law0000150

Validity, Inter-rater Reliability, and Measures of Adaptive Behavior: Concerns Regarding the Probative versus Prejudicial Value

Karen L. Salekin, The University of Alabama

Tess M.S. Neal, Arizona State University

Krystal A. Hedge, Federal Medical Center Devens

Author Note

Portions of this paper were presented at the 2010 annual conference of the American Psychology-Law Society in Vancouver, British Columbia, Canada.

Correspondence concerning this article should be addressed to Karen L. Salekin, The Department of Psychology, 348 Gordon Palmer Hall, The University of Alabama, Tuscaloosa, AL 35487-0348, USA. E-mail: ksalekin@ua.edu

Abstract

The question as to whether the assessment of adaptive behavior (AB) for evaluations of intellectual disability (ID) in the community meet the level of rigor necessary for admissibility in legal cases is addressed. Adaptive behavior measures have made their way into the forensic domain where scientific evidence is put under great scrutiny. Assessment of ID in capital murder proceedings has garnished a lot of attention, but assessments of ID in adult populations also occur with some frequency in the context of other criminal proceedings (e.g., competence to stand trial; competence to waive *Miranda* rights), as well as eligibility for social security disability, social security insurance, Medicaid/Medicare, government housing, and post-secondary transition services. As will be demonstrated, markedly disparate findings between raters can occur on measures of AB even when the assessment is conducted in accordance with standard procedures (i.e., the person was assessed in a community setting, in real time, with multiple appropriate raters, when the person was younger than 18 years of age) and similar disparities can be found in the context of the unorthodox and untested retrospective assessment used in capital proceedings. With full recognition that some level of disparity is to be expected, the level of disparity that can arise when these measures are administered retrospectively calls into question the validity of the results and consequently, their probative value.

Keywords: adaptive behavior measures; *Atkins*; forensic evaluations; admissibility; validity; inter-rater reliability

Validity, Inter-rater Reliability, and Measures of Adaptive Behavior: Concerns Regarding the Probative versus Prejudicial Value

In *Atkins v. Virginia* (2002), the U.S. Supreme banned the execution of individuals with intellectual disability on the basis that doing so would violate the Eighth Amendment's proscription against cruel and unusual punishment. In the 15 years that have passed since the ruling, the assessment of ID has become a central issue in over 436 capital cases, from which only 60 claimants have been found to have ID.[1] (L. Vann, Fellowship Attorney, Death Penalty Resource & Defense Center, personal communication, January 30, 2017). The legal issue in what are now termed "*Atkins* cases," is unlike any other in forensic psychology. In this domain the trier-of-fact is not interested in how symptoms of the condition impact a defendant's ability to participate in judicial proceedings, or if the symptoms present at the time of the crime were sufficient to reduce culpability, the question to be answered is simply whether the condition exists at all.

Intellectual Disability and Adaptive Behavior

There exist two primary definitions of intellectual disability (ID) that, though slightly different, are fundamentally the same. The diagnosis of intellectual disability is made when a person has significant limitations in both intellectual ability and adaptive functioning, with onset occurring during the developmental period (American Association on Intellectual and Developmental Disabilities (AAIDD), 2010; American Psychiatric Association (APA), 2013). According to the AAIDD, adaptive behavior is defined as "the collection of conceptual, social, and practical skills that have been learned and are performed by people in their everyday lives"

[1] This value is derived from a review of decisions reported in the Westlaw electronic legal database and as such, does not include cases that settled at the trial level or that have not been appealed.

(AAIDD, 2010, p. 43)[2] and an almost identical definition is in place for the American Psychiatric Association (APA, 2013). Using the language put forth by the AAIDD, "significant limitations in adaptive behavior ***should (emphasis added)*** be established through the use of standardized measures, normed on the general population, including people with disabilities and people without disabilities" (AAIDD, 2010, p. 43). Significant deficits equate to scores that are "approximately two standard deviations below the mean of either (a) one of the following three types of adaptive behavior: conceptual, social, or practical, or (b) an overall score on a measure of conceptual, social, and practical skills" (AAIDD, 2010, p. 43). Though the AAIDD does not mandate the use of adaptive behavior measures, the use of the word "should" is sufficient to persuade many clinicians and legal professionals that a standardized measure is necessary under most circumstances.

The assessment of intellectual disability dates back to the 1920's, and in search of precision, organizations such as the AAIDD, the APA, the World Health Organization, among others, established ever-changing rules for the assessment of the condition. For more than 30 years, the diagnosis of ID rested solely on measured intelligence, but as of 1961 adaptive behavior was added to the official definition of the American Association on Mental Deficiency (now the American Association on Developmental Disabilities; AAIDD).The first widely-used measure of adaptive behavior was the Vineland Social Maturity Scale (VMS; Doll, 1953), published by the AAMD in 1953. At present, some of the most commonly used measures of adaptive behavior are the Adaptive Behavior Assessment System (currently in its third version; ABAS-III; Harrison & Oakland, 2015), the Scales of Independent Behavior-Revised (SIB-R; Bruininks, Woodcock, Weatherman, & Hill, 1996), and the Vineland Adaptive Behavior Scales

[2] This tri-partite definition of adaptive behavior was adopted by the American Psychiatric Association in the most recent iteration of the Diagnostic and Statistical Manual (DSM-5; APA, 2013).

(currently in its third version; VABS-3; Sparrow, Cicchetti, & Saulnier, 2016). The ratings on these measures are of abilities, and the frequency and level of independence with which they are carried out. All three are used for diagnostic purposes and compare an individual's scores with population norms.

Since *Atkins*, it has become clear that the assessment of ID is anything but straightforward. Though the Justices alluded to the value of the diagnostic criteria set forth by the APA and the AAIDD, the Supreme Court (i.e., the Supreme Court of the United States) left it to the states to develop appropriate ways to enforce the constitutional restriction. However, in a recent decision, the Court curtailed the power afforded to the states and mandated adherence to the well-established medical practice of the applying the standard error of measurement to the interpretation of an intelligence quotient (*Hall v. Florida*, 2014). In *Moore v. Texas*, 2016, the Court further delineated the boundaries of discretion and mandated that adjudications of intellectual disability should be "informed by the views of medical experts." Writing for the majority, Justice Ginsberg reaffirmed that the states do not have unfettered discretion and cannot use non-clinical, judicially-developed criteria for diagnosing intellectual disability. The criteria at issue were the seven factors outlined in *Ex Parte Briseno* (2004), but *Moore* forces all states to determine adaptive functioning consistent with the extant standards of the medical community – not standards promulgated by judges. The reasoning of *Moore* helps to delineate the basis by which the triers-of-fact are to make their determinations regarding adaptive behavior, and may impact the process by which it is assessed.

Unlike the assessment of intellectual functioning, the assessment of adaptive behavior is somewhat unstandardized and subjective, a fact often noted by the judiciary (see for example *Doss v. State*, 2009; *U.S. v. Hardy*, 2010; *Wiley* v. *Epps*, 2010). Though guidelines for

assessment exist, clinicians have much flexibility in choosing the techniques they employ and the weight they place on the information gathered. The type of information gathered, and the weighting thereof, is then used to support their position regarding a claimant's status in relation to ID. Although attempts to standardize the assessment of ID have been made (see AAIDD, 2010; APA, 2013; Schalock et. al., 2010), that which is standard of practice in the community setting (i.e., an assessment conducted in real-time) typically cannot be adhered to when conducting an *Atkins* evaluation. In an *Atkins* hearing, as well as all evaluations conducted in the adult legal system, the assessment is retrospective (Young, Boccaccini, Conroy, & Lawson, 2007).

Case law provides one avenue for review of the methods that forensic clinicians have used to assess intellectual disability and in some cases the triers-of-fact have provided detailed accounts of their formulations of the case and the how they viewed the information provided by the expert(s) (see for example *U.S. v. Smith*, 2011; *Thomas v. Allen*, 2009). In *U.S. v. Smith*, Judge Berrigan noted the following:

> Unlike in a medical, educational, or social services context, the law is concerned with what was rather than what is. The point of an *Atkins* hearing is to determine whether a person was mentally retarded at the time of the crime and therefore ineligible for the death penalty, not whether a person is currently mentally retarded and therefore in need of special services. Because of this, the diagnosis of mental retardation in the *Atkins* context will always be complicated by the problems associated with retrospective diagnosis.
>
> These problems are only compounded by the fact that both the APA and AAMR define mental retardation as a developmental disability and limit the diagnosis to those

> persons who exhibited the required characteristics prior to age 18. As those under the age of 18 are already constitutionally ineligible for the death penalty, *Roper v. Simmons*, 543 U.S. 551 (2005), no clinician evaluating a person for purposes of an *Atkins* hearing will ever be evaluating the person prior to age 18. Mental retardation in the *Atkins* context, if it is to be diagnosed at all, must therefore be diagnosed retrospectively. (p. 43)

Due to the lack of research in the retrospective application of these measures, judges are free to interpret the data how they choose; they may view the standard scores to be valid or they may look at the data in other ways. For example, in *U.S. v. Smith* (2011), Judge Berrigan went to great lengths to evaluate the consistency of ratings and, due to vastly different opinions of opposing experts, she chose to go beyond review of standard scores and conducted an evaluation of differences at the item level. Specifically, she compared individual scores, for each question, for three raters on the Vineland Adaptive Behavior Scale – Second Edition (VASB-II; Sparrow, Cichetti, & Bella; 2005; 429 items) and two raters on the Adaptive Behavior Assessment System (ABAS-II; Harrison and Oakland, 2003; 239 items). In *Smith*, the expert for the State was of the opinion that the raters deliberately lowered their ratings to ensure that the deficits in adaptive behavior would meet threshold; the expert for the claimant was not.

The results of Judge Berrigan's investigation of the VABS-II resulted in perfect consistency for 88% of the items, and 77% and 88% when two of the three raters (i.e., mother, an older sister, and a younger sister) produced identical scores. She noted that when discrepancies were found, the majority of scores were within one point of each other; she further noted that two point discrepancies occurred only 8% of the time. As noted by Judge Berrigan, "This consistency strongly supports the reliability of the tests and the conclusion that the three respondents were not deliberately exaggerating his deficits" (p. 55). Of interest, consistency on the ABAS-II was

substantially lower with approximately 54% of the answers between Smith's sisters having been identical, 43% with one level of disparity, and 3% with more than a one level difference. Judge Berrigan made no comment on her views of these findings.

As evident by the above example, the use of adaptive behavior measures in *Atkins* hearings is less than perfect and conclusions arise on the basis of a combination of clinical judgment, common sense, and strategies not yet examined. The question for the Court is whether adaptive behavior measures are sufficiently reliable to be admitted into legal proceedings when used in an unorthodox, untested manner, with no known rate of error, and in a manner that is not generally accepted by the scientific community. Another way to conceptualize the retrospective evaluation is to ask whether a person's current memory, of their past perceptions, of another person's behavior, at a specific point in the distant past, in any way comports with reality.

The *Daubert* Trilogy

Problems associated with the assessment of vaguely defined constructs, as is the construct of adaptive behavior, are not new and neither are the associated issues of admissibility. One need only look to the history of admissibility of expert testimony to know that the judiciary has been struggling to find a balance between acceptance of new (and at times old) science and permitting only that which meets the indiscernible line of acceptability. An in-depth discussion of the rules of admissibly is outside of the scope of this paper, but understanding the trajectory from early decisions, such as *Frye v. United States* (1923) to present, is important.

Daubert v. Merrell Dow Pharmaceuticals (1993), *General Electric Co. v. Joiner* (1997), and *Kumho Tire Co. v. Carmichael* (1999) together constitute what is oft referred to as "The *Daubert* Trilogy." Arising from this trilogy are guidelines and procedures for determining the evidentiary reliability of expert testimony and subsequent admissibility (Merlino et. al, 2007).

The trilogy extended the analysis from legal merits and general acceptance (*Frye v. United States,* 1923) to include judicial scrutiny regarding qualifications of the experts, their methods of investigation, and the conclusions drawn from those procedures (Merlino et. al, 2007). As per *Daubert*, expert testimony is evaluated on the following five factors:

> (1) whether the expert's technique or theory can be or has been tested—that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability;
>
> (2) whether the technique or theory has been subject to peer review and publication;
>
> (3) the known or potential rate of error of the technique or theory when applied;
>
> (4) the existence and maintenance of standards and controls; and
>
> (5) whether the technique or theory has been generally accepted in the scientific community.

With regard to the analysis, the *Daubert* Court stated that the inquiry is flexible and the focus must be on the relevance and reliability of the expert's methodology, rather than the conclusions generated by those methods. Hence, these factors are neither exhaustive nor dispositive.

In *General Electric v. Joiner* (1997), the Supreme Court noted that the essence of *Daubert* is to ensure that the evidence admitted is not only relevant but also reliable, and the Justices stressed that the link between the scientific testimony and facts must be directly applicable to the case at hand. Echoing the decision in *Daubert*, the *Joiner* court noted that trier-of-fact must be able to evaluate the relevance and reliability of the experts' methodology – not just their conclusions.

Kumho Tire Co. v. Carmichael (1999) dealt with the type of evidence that fell under the gatekeeping role of the Court. The Court ruled that the gatekeeping obligation applies not only to scientific testimony, but to all expert testimony including that which is technical or otherwise specialized. Furthermore, the Court reiterated the flexibility of the gatekeeping criteria, holding that judges may consider one or more of the factors articulated in *Daubert* when determining reliability, but that those factors need not necessarily nor exclusively be applied to all experts, or in every case. The Court concluded that,

> the trial court must have the same kind of latitude in deciding *how* to test an expert's reliability, and to decide whether or when a special briefing or other proceedings are needed to investigate reliability, as it enjoys when it decides *whether* that expert's relevant testimony is reliable. (p. 152, emphases in original)

The Federal Rules of Evidence 702 (F.R.E. 702) clarifies when and how witnesses are qualified to testify as experts. This rule, first put forth in 1973 and modified in 2000 and 2011, was referenced numerous times in the *Daubert* trilogy. As per Rule 702,

> A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:
>
> (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
>
> (b) the testimony is based on sufficient facts or data;
>
> (c) the testimony is the product of reliable principles and methods; and
>
> (d) the expert has reliably applied the principles and methods to the facts of the case.

Upon review of the trilogy and the F.R.E., it is clear that findings based on the retrospective use of adaptive behavior measures[3] may not meet the standards of admissibility.

Measurement

From 1932 to 1940, the British Association for the Advancement of Science debated the meaning of measurement, and after much discussion, the 19-member committee came to some agreement that the broadest and most useful definition of measurement is "the assignment of numerals to things so as to represent facts and conventions about them" (Stevens, 1946; p. 680). As noted by Stevens (1946), what is and is not measurement boils down to the answer to one question: "What are the rules, if any, under which numerals are assigned? If we can point to a consistent set of rules, we are obviously concerned with measurement …."(p. 680).

Measurement is the foundation of psychological testing. A psychological test provides a systematic method for obtaining one or more samples of behavior and for scoring and evaluating those samples according to empirically derived standards (Anastasi & Urbina, 1997; Urbina 2004). Though psychological tests are only one component of an assessment, the results of testing are used to make important decisions. In her discussion of psychological tests as tools, Urbina (2004) discussed the value of tests, but added a cautionary note: "Like other tools, psychological tests can be exceedingly helpful – even irreplaceable – when used appropriately and skillfully. However, tests can also be misused in ways that may limit or thwart their usefulness and, at times, even result in harmful consequences" (p. 4). Part of ensuring that the results of testing are not misused is to ensure that the test are administered, scored, and interpreted in the manner with which they were developed. Once again, when used to

[3] The points of analysis in this manuscript focus solely on adaptive behavior measures, not the assessment of adaptive behavior in an *Atkins* case. Adaptive behavior measures, when used in such cases, represent one method of assessing adaptive behavior among many that must be considered by the evaluating clinician and the presiding judge.

retrospectively evaluate adaptive behavior, measures of adaptive behavior fall short of the ideals of measurement as the retrospective administration, scoring, and interpretation of the data is not consistent with the manner in which the adaptive behavior measures were developed.

Measures of Adaptive Behavior: Test Administration and Interpretation

Respondents

By design, measures of adaptive behavior are to be completed by knowledgeable respondents, based on their recent observations of an individual's behavior; in other words, these measures are completed in real time, not at some time in the distant past. In the manual, the authors of the ABAS-3 stress the importance of the type of respondent, their knowledge of the person, and their current level of contact (Harrison & Oakland, 2015):

> All respondents should have had frequent, recent, prolonged contact with the individual (e.g., most days, over the last few months, for several hours each day). These contacts must have offered the respondent an opportunity to observe the various adaptive skill areas measured by the ABAS-3. (p. 9)

In only the rarest of circumstances can these rules be adhered to, so from the outset the validity of the results are almost always called into question.

Multiple Domains

Assessing typical performance across multiple settings is necessary because reliance on an individual's functioning in one setting may provide an inaccurate portrayal of an individual's ability to function on a day-to-day basis (AAIDD, 2010; Greenspan & Switzky, 2006; Harrison & Oakland, 2015; Macvaugh & Cunningham, 2009; Sparrow et al., 2005; Widaman & Siperstein, 2009). For children and youth, information regarding functional ability within both the home environment and the community can almost always be obtained. For the most part,

assessments of children and youth will also include ratings from informants who are also knowledgeable about their behavior within academic settings and these recollections and observations can be supplemented by records that are available and being created in present time (e.g., special education; resource education; mainstream education; vocational training).

The type of data available for children and youth is not always available for evaluations of adults. For many *Atkins* claimants, the evaluations are conducted many years after they exited the developmental period and the passage of time negatively impacts the collection of data. Many times collateral sources from the community (e.g., neighbors; friends; employers; store clerks) cannot be located and even if located, their memories are often poor and are always influenced by experiences that have occurred since the individual turned 18. Teachers from the past are often deceased or cannot be found, and those found have memories that have changed with the passage of time and experiences. In addition, academic records are often unavailable because they have been destroyed or the data is insufficient to support or refute deficits in this domain.

Multiple Raters

In addition to selecting appropriate raters across multiple domains, it is standard practice to obtain data from multiple raters (AAIDD, 2010). The underlying rationale is that the assessment will be more accurate when evaluating behavior across informants and across domains (see for example AAIDD, 2010; Harrison & Oakland, 2015; Macvaugh & Cunningham, 2009; Widaman & Siperstein, 2009). As noted by Harrison and Oakland (2003), “the use of multiple respondents can provide information about the degree of consistency of an individual’s adaptive skills across settings, in response to different environmental demands, and from the unique perspectives of different respondents” (Harrison & Oakland, 2003, p. 19).

Inter-rater Reliability

While it is true that consistency allows for a higher level of certainty that the information accurately reflects an individual's abilities, the opposite may not be true (see for example, Achenbach, McConaughy, & Howell, 1987; Szatmari, P., Archer, L., Fisman, S., & Streiner, D. L., 1994; De Los Reyes & Kazdin, 2005; De Los Reyes, A., 2011). This sentiment was clearly articulated by Voelker, Shore, Hakim-Larson and Bruner (1997) two decades ago in the context of behavioral ratings of children:

> Obtaining reports from informants who know the child in different contexts, such as the child's parent and teacher, increases the behavioral repertoire sampled and provides a more complete description of the child's skills. However, this breadth of behavioral sampling can result in low rates of agreement between informants, raising questions about the source of the inconsistency. Discrepancies between reports of teachers and parents may reflect unreliability or lack of comparability of the measure(s), rater bias, or genuine differences in the child's behavior between the two environments. Evaluators are often encouraged to take advantage of multiple sources of information for making decisions regarding program planning for children (e.g., Sattler, 1988), but this advice presumes both good interrater reliability for the measure(s) used and a sufficient research base to permit evaluation of any systematic sources of disagreements that occur. (p. not provided)

The research that has identified inconsistency as the norm is largely from child and adolescent studies of mental health rather than those of adaptive behavior. The lower level of consistency may be due to the fact that some of ratings on these scales require memories and speculation rather than observation. However, there exists research in the area of functional behavior

analysis (Newton & Sturmey, 1991; Paclawskyj et al., 2001; Sigafoos, Kerr, & Roberts, 1994; Thompson & Emerson, 1995; Shogren & Rojahn, 2003; Zarcone, Rodgers, Iwata, Rourke, & Dorsey, 1991) that supports the position that the findings of this research transfer to ratings of adaptive behavior.

With the acknowledgment that perfect concordance is not expected, test developers provide inter-rater reliability statistics within and across domains. The fact that these studies are conducted implies that some level of concordance is expected and that there is a point where concordance is deemed unacceptable. Generally speaking, reliability coefficients reflect the proportion of "true" information about a construct in comparison to that random variability. For example, if inter-rater reliability was found to be .8, 80% of the variability in scores reflects a measure of the construct and 20% something else. The higher the reliability coefficient, the more certain one can be in that the measure is tapping a construct that can be measured.

Test developers almost invariably offer inter-rater reliability coefficients for two raters from the same setting (e.g., family = two parents; academic = two teachers/teacher's aides), which is necessary and appropriate. However, as is evident below, the number of studies conducted by the developers of the SIB-R, the ABAS (versions II and III), and the VABS (versions II and 3) are few and the samples less than optimal for evaluating consistency nearing the tails of the distribution. There are only three studies conducted with individuals with intellectual disability, all of which are problematic (Bruininks, Woodcock, Weatherman, & Hill, 1984; Bruininks, Woodcock, Weatherman, & Hill, 1996). First, they were done many years ago, and second, the samples are not reflective of those who enter the legal system. One study utilized a sample of children diagnosed with moderate intellectual disability and the others utilized samples of children and/or youth diagnosed with either the moderate and/or severe intellectual

disability. The findings of these studies, while important, are of little value an assessment of adaptive behavior within the legal system because most individuals fall in the mild category of intellectual disability.

Inter-rater Reliability: SIB-R, ABAS II and III, and VABS II and 3

Scales of Independent Behavior-Revised (SIB-R; Bruininks, Woodcock, Weatherman, & Hill, 1996). As part of the norming process, the authors of the SIB-R conducted three very small-scale studies of inter-rater reliability with children and raters from two domains (home; school) who had opportunities to observe the child in the same setting. The sample size for study one was 26 and the raters were fathers and mothers. The sample size for study two was 30 and the raters were teachers and teacher's aids. Correlations were high and ranged from .88 to .97 for children with moderate intellectual disability; similar correlations were found with a sample of typically-developing children.

In addition to the above-noted studies, the authors of the SIB-R included inter-rater reliability data from the original version of the measure (SIB; Bruininks, Woodcock, Weatherman, & Hill, 1984). The authors evaluated the reliability between teachers and teacher's aids and the results demonstrated moderate to high correspondence at .80 for the Broad Independence Score, and the cluster scores ranged from a low of .74 to a high of .86. The sample used for this study included 39 adolescents with moderate or severe intellectual disability.

Adaptive Behavior Assessment System – Second Edition (ABAS-II; Harrison & Oakland, 2003). The authors of the ABAS-II conducted four interrater reliability studies (N = 309). Of the four studies, one was conducted using the Teacher/Daycare rating forms, two with the Teacher rating forms and one with the Parent/Primary Caregiver forms. All studies were

conducted using samples of typically-developing children and youth. Results from these limited studies showed moderate to high consistency (correlations ranged from .58 to .93).

- Teacher/Daycare (N = 42): ages 2-5 years; Corrected *r* for the Conceptual Domain = .83; Social Domain = .74; Practical Domain = .74; Global Adaptive Composite = .83
- Teacher (N = 34): ages 5-9 years; Corrected *r* for the Conceptual Domain = .58; Social Domain = .74; Practical Domain = .92; Global Adaptive Composite = .93
- Teacher (N = 50): ages 10-18 years; Corrected *r* for the Conceptual Domain = .72; Social Domain = .75; Practical Domain = .88; Global Adaptive Composite = .90
- Parent/Primary Caregiver (N = 56): ages 0-5 years; Corrected *r* for the Conceptual Domain = .86; Social Domain = .72; Practical Domain = .77; Global Adaptive Composite = .82

Adaptive Behavior Assessment System – Third Edition (ABAS-3; Harrison & Oakland, 2015). Interrater reliability studies conducted for the ABAS-3 included the Parent/Primary Caregiver Form (two samples delineated by age categories), the Teacher/Daycare Provider Form (two samples delineated by age categories), and the Adult Form Rated by Others[4]. Of the five studies, the results of four are applicable to individuals from birth to age 21, while the fifth was conducted using the Adult Form, Rated by Others (applicable to people age 16 to 89). The correlations indicated moderate to strong correspondence between raters on all forms (correlations ranged from .68 to .92); as would be expected, the General Adaptive Composite (GAC) scale demonstrated the highest correspondence.

[4] Information regarding the raters is not provided in the manual for the ABAS-3.

According to the authors, "the results show that the interrater reliability of the ABAS-3 scores is acceptable for clinical use and consistent with that of other behavior rating scales"(p. 80). Based on a notation in the manual, it appears that the normative samples did not include individuals diagnosed with intellectual disability:

> … sampling methods are designed to include cases with mild disabilities, as long as the severity does not preclude mainstream activities (such as general education). Over a large standardization sample, these methods are designed to include these mild problems at their population base rate. (p. 60)

As is true for other measures, this exclusion limits the applicability of the measure in evaluations conducted within the legal system where the evaluation and diagnosis of intellectual disability is the central issue.

Vineland Adaptive Behavior Scales – Second Edition (VABS-II; 2008). Inter-rater reliability studies for the VABS-II were provided for the Parent/Caregiver Rating Form (appropriate for individual's birth to 90) and the Survey Interview Form (appropriate for individual's birth to 90). Both studies utilized samples of typically-developing individuals. The VABS-II provides scores for five sub-domains (i.e., Communication; Daily Living; Socialization; Play and Leisure; Motor) and the Adaptive Behavior Composite (ABC). Using a sample of 112 typically developing individuals (ages birth to 18 years) the authors found moderate concordance for the sub-domains (range .72 to .80), as well as the ABC (i.e., .78). Similar results were found for the Parent/Care Giver Form (N = 152; ages birth to 18 years) concordance for the sub-domains ranging from .71 to .83 and an ABC of .82. .82. Inter-rater reliability statistics were not provided for the Parent/Caregiver Form as applied to a sample of typically developing individual's age 19-61 (n=39), and as noted in the manual, this was because

inter-rater reliabilities "were strongly affected by the large proportion of cases scoring at or just below the maximum score on subdomains"(p. 119).

Vineland Adaptive Behavior Scales –Third Edition (VABS-3; Sparrow, Cicchetti, & Saulnier, 2016). Three inter-rater reliability studies were conducted, all with individuals that fall in the age category of birth to age 20 years (i.e., Parent/Caregiver; Teacher/Teacher; Interviewer/Interviewer). Correlations were noted to be good to excellent range (e.g., corrected *r* ranged from .70 to .81 on the Comprehensive Form Domains of the Interview Form). According to the authors, the results of the inter-rater reliability conducted with adults were uninterpretable due to the ceiling effect of the sample.

The normative sample included individuals with developmental disabilities and did so based on the U.S. government statistics on special education services (National Center for Education Statistics, 2014). As noted in the manual, sampling targets included individuals with intellectual disability, developmental delay, autism, emotional disturbance, specific learning disability, and speech/language impairment; other classifications were grouped together in an "other IDEA disabilities categories." The percentage of individuals in the normative samples for the three forms of the VABS-3 (i.e., interview; parent/caregiver; teacher) varied, but were reported to closely match that of the population of the United States. Given the method used to norm this measure, there is no information regarding inter-rater reliability for individuals with intellectual disability.

Memory

All information is subject to distortion and the data obtained is inextricably linked to the accuracy of memory and the accuracy of perceptions at the time of contact; one of which is known to be fallible (memory; see for example, Loftus, 2003) and the other known to be affected

by multiple factors unique to the perceiver and each situation (perceptions; De Los Reyes & Kazdin, 2005). Thorough coverage of this topic is beyond the scope of this paper, and as such, what is mentioned here is that deemed relevant to understanding memory in the context of completing measures of adaptive behavior in a retrospective manner.

While atypical, retrospective accounts of the behavior and abilities of others can be viewed as autobiographical memories. Raters must think about a specific time-period in their lives and bring forth a memory of having witnessed another person carrying out a task (or not), how well and/or how often they carried out that task, and whether they carried out that task with or without assistance. Autobiographical memories are reconstructions of the past that are prone to error and distortion, and are influenced by knowledge, attitudes and beliefs (Schacter, 2012). Hence, the data obtained from retrospective administration of measures of adaptive behavior cannot reflect reality, but instead reflect the perceptions of the rater gleaned from a reconstruction of their past.

Schema Theory

Research has shown that memories are organized and stored in a manner that assigns meaning to experience (Bartlett, 1932); memories guide behavior (Bartlett, 1932; Shea, Krug, & Tobler, 2008; Kumeran et al., 2009), and facilitate encoding and retrieval (Anderson, 1984; Preston & Eichenbaum, 2013; Preston & Eichenbaum, 2013; Van Kesteren et al., 2013). One way researchers and theorists have conceptualized the organization and storage of memories is based in schema theory (Bartlett, 1932; Markus & Zajonc, 1985). At the most basic level, schemas provide a way to conceptualize how knowledge obtained from prior experiences influence an individual's attention and behavior and their interpretation of events, and as such, influence the reconstruction of memories (Lampinen, Copeland, & Neuschatz, 2001).

The ratings of prior behavior, and the standard scores produced therefrom, are the product of the amalgamation of data that has been collected and integrated on a daily basis for many years. By the time an individual has reached adulthood, many actions have been frequently expressed and are typical behaviors associated with everyday living; this is true for individuals diagnosed with ID and those without, though the range of behavior is not the same. Research has shown that when asked to recall an event that is an example of an action or experience that has been repeated many times (e.g., puts dishes away), semantic memories are produced. A semantic memory is a memory of a prototypical experience that has occurred many times and when accessed the outcome is merely an inference that may or may not reflect reality (Belli, 1988; Brewer, 1986; Jobe et al., 1990; Means & Loftus, 1991; Menon, 1994).

Memory reconstruction and admissibility

With recognition of the fallibility of memory, the British Psychological Society generated a set of guidelines regarding memory processes, as the findings might be applied within the legal arena (Conway & Holmes, 2008). As would be expected, the overarching theme was that human memory is fallible and much of the time it is difficult, if not impossible, to tease out reality from reconstruction. The research group identified 10 key points, five of which are relevant to memory in the context of the retrospective assessment of adaptive behavior:

1. Memories are records of people's experiences of events and are not a record of the events themselves.
2. Remembering is a constructive process.
3. Memories typically contain only a few highly specific details.
4. The content of memories arises from an individual's comprehension of an experience, both conscious and non-conscious.

5. People can remember events that they have not experienced.

The authors of the guidelines discussed the implications of memory research in the context of the typical legal proceedings where the issue arises (e.g., eye-witness testimony; memory and stress; witness interviews) and not within the context of the *Atkins* proceedings. However, as previously mentioned, the accuracy of memory is paramount in *Atkins* proceedings as they include data derived from memory (e.g., standard scores from measures of adaptive behavior) and these data are entered into evidence and conceptualized as accurate representations of the past. Review of the structure of these measures immediately calls into question the accuracy of the data produced, as well as their admissibility in legal proceedings.

All measures of adaptive behavior have methods by which information obtained by raters is represented by one or more statistics that provide a meaningful representation of a construct (e.g., standard scores; rankings; age equivalents). Though rating options and instructions differ by measure, raters must have very specific knowledge of a person's ability to carry out many tasks and how often they do them without the assistance of others. For example, the Vineland-3 Comprehensive Interview Form consists of 458 items across 11 adaptive behavior subdomains, with three rating options and a fourth that permits the rater to provide an estimate rather than a known. The SIB-R is similarly detailed with 315 questions spread across 14 sub-scales with four rating options and the ability to identify when responses reflects guesses rather than knowledge. Unlike the Vineland–3, the SIB-R places contingencies on the ratings that reflect how well the individual carries out the task, the percentage of time they carry out the task and whether they need to be asked to do so. As evident above, completing these instruments in accordance with how they were designed is difficult and at times requires making an inference; to think that the statistical data obtained from retrospective accounts represent "ground truth" is unrealistic.

Inter-Rater Reliability, Intellectual Disability, and the Retrospective Assessment of Adaptive Behavior

It has been 15 years since the ruling in *Atkins v. Virginia* (2002), but few studies have been conducted in that time on the retrospective assessment of intellectual disability as relevant to legal cases. For instance, Doane and Salekin (2009) examined how susceptible the ABAS-II and SIB-R are to feigned adaptive behavior deficits, finding that faking was easily detected by the SIB-R, but not by the ABAS-II. More recently, Boccaccini, Kan, Rufino, Noland, Young-Lundquist, and Canales (2016) examined correspondence between correctional staff ratings and offender ratings of adaptive behavior, using the ABAS-II. They found that correctional staff assigned significantly lower scores than probationers, with other findings suggesting the offenders' self-report scores were likely more valid than the correctional officers' reports. Their findings highlight the limitations of using correctional staff members as respondents for adaptive behavior assessments.

Though few studies, like those mentioned above, have been conducted in this area since the *Atkins* ruling, not a single empirical study has been conducted on the inter-rater reliability of measures of adaptive behavior when used with individuals with mild ID, nor has there been a study on the inter-rater reliability of retrospective assessment. In order for triers-of-fact to evaluate this data in line with the holding in *Daubert* and the F.R.E. 702, research in both of these areas is necessary. At the time of writing, there is no way to know the level of concordance to expect under the best circumstances for assessments of individuals with mild ID (i.e., ratings made in real-time during the developmental period), never mind those conducted in the legal system (i.e., retrospective).

Three case examples are presented, each of which illustrate how measures of adaptive behavior can interfere with diagnostic and legal decision-making. One case was conducted in juvenile court and as such, was conducted in real-time and within the developmental period. The other two cases were conducted retrospectively; one close in time to the developmental period and the other well outside of this time-period. Together, these cases show the lack of concordance that can occur with the results so disparate that it is difficult to know how to use the information, or if the data should be used at all.

Case 1: Youthful Offender

To demonstrate that high variability can be obtained on measures of adaptive behavior, even when assessments are carried out in accordance with the standards of practice, a case from juvenile court is provided. In this case, the clinician was asked to conduct an aid to sentencing evaluation for a youth with no previous criminal history who had been adjudicated delinquent secondary to the commission of a crime that involved the use of a firearm. Since the defendant was a juvenile (i.e., 16 years of age) there was no need to conduct a retrospective assessment of adaptive behavior. The youth's Full Scale IQ was measured to be 67 with the Wechsler Adult Intelligence Scale– Fourth Edition (WAIS-IV; Wechsler, 2008), and his age-based standard score in Total Achievement on the Woodcock-Johnson Tests of Achievement – Third Edition (WCJ-III; Woodcock, McGrew, & Mather, 2001) was 65 (1^{st} percentile), indicating he was performing well below his same-age peers in the academic setting.

At the time of the evaluation, this youth was living with his aunt and uncle and had done so on an intermittent basis for several years. His placement in his aunt's home was due to difficulties within his immediate family, but despite family discord, this youth had frequent and lengthy contact with his mother.

The ABAS-II was completed by his mother, his maternal aunt, and three of his teachers (two of his general education teachers and his special education tutor). All assessments were contemporary in time and raters appropriate with regard to level of knowledge and type and extent of contact. In this case, the judge disregarded the findings of all adaptive behavior measures. Figure 1 and Table 1 show the results of the ABAS-II and demonstrate the lack of concordance that can occur, even when the assessment is carried out in real time.

[Insert Figure 1 and Table 1 here]

Case 2: Results from an assessment of ID in the context of *Atkins*

The second case example shows data that was obtained for a trial level *Atkins* proceeding; the assessment was conducted seven months after the defendant's incarceration on the charge of capital murder. At the time of the crime the defendant was 27 years of age. Three people, all family members, completed the SIB-R and did so based on their recollections of the defendant's functioning 11 years prior (i.e., when he was 16 years of age). This defendant had been assessed for intellectual disability in the past and his full-scale IQ scores ranged from 53 to 62. Based on current testing, his full scale IQ was measured to be 51 and his grade-based standard score in Total Achievement on the Woodcock-Johnson Tests of Achievement – Third Edition (WCJ-III; Woodcock, McGrew, & Mather, 2001) was 60 (grade equivalent = 2.2) indicating he was performing well below his same-age peers in the academic setting.

The defendant's mother was one of the raters and, with the exception of his incarceration, she had been in frequent and daily contact with her son since his birth. At the time she completed the ratings, her contact was limited to once per week, for brief periods of time. His younger sister was chosen as a second rater; her contact with her brother up to the age of 18 years was daily and for extended periods of time. She had moved out of the family home two years prior to his

incarceration, but during that time she frequently visited with him. Most recently her contact was limited to approximately once per month during visitation at the jail. The third rater was his maternal aunt, who had known the defendant since birth. She had always been in frequent, though not daily, contact with him when he lived in the community. In addition to contact with the defendant, she was in frequent contact with his mother via telephone and in-person visits; these contacts continued following the defendant's arrest and incarceration. Since his arrest, his aunt had only seen him during court proceedings. Figure 2 and Table 2 show the results of the SIB-R and once again, demonstrate the lack of concordance between raters; this time for the retrospective assessment of adaptive behavior.

[Insert Figure 2 and Table 2 here]

Case 3: Results from an assessment of ID in the context of *Atkins*

The third case example provides the data obtained from another assessment conducted for an *Atkins* hearing at the trial level. The defendant was 20 years of age at the time of the evaluation and due to his arrest (at age 19 years) and subsequent incarceration, approximately 11 months had passed since each rater had been in contact with him in a community setting, and 24 months since he exited the developmental period (i.e., passage of time between assessment and recollection was approximately two years). Between the time of his arrest and the time of the assessment, visitation between raters and the defendant was frequent, though for only short periods of time. All raters were family members, two sisters and one brother, and deemed appropriate in terms of relationship and prior level of contact (i.e., daily prior to his arrest).

Despite the assessment having been conducted shortly after his transition into adulthood, raters outside of the family were not readily available. Three teachers and two employers were approached and all provided anecdotal information via interview; none of these individuals

believed that their memories could produce accurate ratings, nor did they believe that they had enough information to complete such an extensive behavior rating scale. In this case friends were few and work history limited to three jobs, all of which required limited skills (e.g., sweeping floors; picking up trash strewn on city property).

As is evident from Figure 3 and Table 3, concordance was much higher in this case than those previously presented and in the opinion of the evaluator, was due to the lower level of functioning demonstrated by this individual in the community setting, as well as within the jail. The defendant demonstrated deficits in multiple domains that were easily seen and frequently expressed. These deficits impacted his ability to function in domains including, but not limited to, work, school, self-care, inter-personal relationships, and personal safety. Unlike many cases, information regarding his functioning was corroborated by multiple sources including anecdotal information provided by many third-party informants (e.g., teachers; employers; friends; family members) and records. Regarding measured intelligence, at the age of 18 years his full scale IQ was measured to be 57 and his IQ as per the current evaluation was found to be 67.

[Insert Figure 3 and Table 3 about here]

Discussion

Given the gravity of legal proceedings, it would be expected that the tools used to assess status in relation to ID would be much more precise than has been demonstrated in the case studies provided. If these measures are tapping an underlying construct that can be measured and used as evidence to either support or refute the diagnosis of intellectual disability, there needs to be guidance as to how to interpret conflicting data. Should ratings only be taken from one source? If so, which source should be considered to be best? Is it permissible to take some scores

from one rater and some from another? Can one discount the perceptions of a single rater if they are out of line with others?

In a community setting, the questions outlined above typically wouldn't be asked, and the use of clinical judgment with the subjective meshing together of data would not raise concern. However, in legal contexts, these issues are of significant concern. The legal setting is adversarial and there is disagreement among evaluators the trier-of-fact must sort out the real from the simulated, the truth from the partial truth, and that which is believed to be true, but is not.

Both the AAIDD and the APA have created a demarcation between what qualifies for a diagnosis of ID when AB measures are used and what does not, at least when the measures are used in accordance with standard of practice. As previously noted, the AAIDD defines significant deficits in adaptive behavior in the following manner: significant deficits equate to scores that are "approximately two standard deviations below the mean of either (a) one of the following three types of adaptive behavior: conceptual, social, or practical, or (b) an overall score on a measure of conceptual, social, and practical skills" (AAIDD, 2010, p. 43).

Scores obtained from measures of adaptive behavior are purported to reflect something within the person that can be measured and quantified, specifically an individual's adaptive functioning. However, these cases demonstrate such a lack of concordance that the data obtained from the retrospective application cannot be reflective of a single construct that exists in one form that can be measured and numerically expressed. Perhaps the scores obtained should not be considered reflective of a person's adaptive behavior, but within the context of a retrospective assessment, should be re-conceptualized as a person's memory, of their prior perception, of another person's functional abilities, at some point in the distant past. In this way, the trier-of-

fact is no longer looking for the "truth" when interpreting the scores, but rather a way to compare that person's perspective, as depicted by standard scores, to a normative sample. Comparing scores produced by raters is necessary, but considering one to be "real" and another to be biased, inaccurate, or otherwise tainted may be unwarranted.

The findings of research in the field of memory can assist clinicians and triers-of-fact with developing a strategy to evaluate data and make decisions about what to do with it. Recognition of the fact that human memory is not a recording of events, but is instead a reconstruction of one or more events, that may or may not have happened, that lacks detail or the detail provided is in error, is important for both the clinician and the trier-of-fact. Within this framework, interpretation of standard scores shifts from the focus from the claimant to the rater.

As noted by De Los Reyes (2011), "support for informant discrepancy as a substantive construct comes from decades of basic psychological research in interpersonal perception and memory recall that broadly focuses on how different people often have different views of the same people or sets of behaviors" (p. 3). De Los Reyes was referencing multiple informant discrepancies when administered correctly, but clearly this statement holds true for retrospective assessments. In a 2005 manuscript, De Los Reyes and Kazdin explained informant discrepancies within the framework of the Attribution Bias Context Model. According to this model, discrepancies exist because informants differ along the following dimensions: (a) Attributions of cause (i.e., dispositional qualities vs. environmental constraints), (b) the biases or decision thresholds that guide decision-making, and (c) the contexts within which informants observe the behavior (e.g., home, school). These dimensions, understood in the past and the present, permit the understanding of data produced by measures of adaptive behavior.

There is a clear need for research that examines the validity of information obtained via adaptive behavior measures when administered retrospectively. The best data would come from studies that compare scores on a measure of adaptive behavior previously completed in accordance with the rules of administration with one completed years later. While this option is viable, studies have not yet conducted. A second method of assessing validity, though not ideal, is by looking at inter-rater reliability when an assessment is conducted with multiple raters, on the same adaptive behavior measure, for the same period of time, and completed at the same time. If concordance is high, one can infer that the scores represent reality, at least to some degree (see for example, Achenbach, McConaughy, & Howell, 1987; Szatmari, P., Archer, L., Fisman, S., & Streiner, D. L., 1994; De Los Reyes & Kazdin, 2005; De Los Reyes, A., 2011).

In addition to sound research on how these tools perform retrospectively and with multiple informants, research is also needed on the use of these tools in correctional and forensic contexts. Emerging research suggests that correctional staff members may not be a valid source of information for adaptive behavior assessments, as data show correctional staff may systematically underestimate adaptive behavior abilities in offenders (Boccaccini et al., 2016). This has been an issue that has been addressed by the AAIDD (2010; Schalock, 2010) and the association has been clear in stating their disapproval of this use of adaptive behavior measures.

As per *Daubert*, an expert's theory or technique must have been tested, subject to peer review, have a known or potential rate of error, have and maintain standards and controls, and be generally accepted in the scientific community. If the court were to consider the typical assessment of adaptive behavior – that is, inter-rater reliability conducted in real-time, with typically-developing individuals (primarily children), in one domain - it is likely the *Daubert* criteria would be met. However, what is typical in community settings is not what is typical in

the context of adult legal proceedings. In legal contexts, perhaps none of the admissibility criteria would be met for the retrospective administration of adaptive behavior measures.

The results of these case studies do not support the admissibility of the results from adaptive behavior measures - at least not if the results are supposed to meet the standards put forth by *Daubert* or the F.R.E. and they are considered to be measures of adaptive behavior. Given the flexibility in *Daubert* and that admissibility is the purview of only the judge**,** the court may to choose to admit the results of the adaptive behavior measures administered retrospectively. That said, unlike much data provided to the court, there is no research to support the validity of results from adaptive behavior measures that were administered retrospectively.

The problems associated with the retrospective use of adaptive behavior measures are part of a broader issue that pertains with the methods by which clinicians can, or perhaps should, conduct a retrospective assessment of ID. This calls for research that has the potential to improve accuracy in this area of psychological assessment and the application of the data to this area of the law. Inter-rater reliability (as was the focus of this paper), test-retest reliability (pre-18 years versus post 18 - multiple raters from different contexts), inter-rater reliability (claimant versus rater), and the validity of cross-cultural application, are among many areas of needed research.

It will take time for research to catch up with the needs of the legal system, and meanwhile the courts will continue to have to make determinations regarding who has intellectual disability and is thus not eligible for the death penalty. Clinicians will need to decide whether to use adaptive behavior measures or not, and to explain that decision in a court of law. These decisions should be guided by research (e.g., memory; typical assessment of ID), best practice, and ethical guidelines. The courts must be informed of the complexity of an assessment

of intellectual disability in the adult criminal justice system so they can ensure that they have the requisite knowledge and understanding of the disorder and the methods by which it is assessed. This knowledge and understanding will give the court the basis to make informed determinations. In the end, if admitted or not, the data produced from the retrospective administration of adaptive measures is inherently flawed and cannot be construed as a measure of adaptive behavior. Once again, the data must be recognized for what they are – simply numbers that represent a person's current memory of their past perceptions, of another person's behavior at a specific point in the distant past, the accuracy of which is unknown.

References

Achenbach, T. M. (2006). *As others see us: Clinical and research implications of cross-informant correlations for psychopathology*. Current Directions in Psychological Science, 15, 94–98.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child-adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin,* 101, 213–232.

American Association on Intellectual and Developmental Disabilities (AAIDD; 2010). *Intellectual disability: definition, classification, and systems of supports (11th ed.).* Washington, DC: AAIDD.

American Psychiatric Association. (APA; 2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

American Psychological Association, Office of General Counsel (2017, January 23). *Moore v. Texas*. Retrieved from: http://www.apa.org/about/offices/ogc/amicus/moore.aspx

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Anderson, R. C. (1984). Role of the reader's schema in comprehension, learning, and memory. In: R. Anderson, J. Osborn, & R. Tierney (Eds.), *Theoretical models and processes of reading* (4th ed.). Newark: International Reading Association.

Atkins v. Virginia, 536 U.S. 304 (2002).

Bartlett, F. C. (1932). Remembering: A study in experimental and social psychology. Cambridge, England: Cambridge University Press.

Belli, R. F. (1988). Color blend retrievals: Compromise memories or deliberate compromise responses? *Memory and Cognition, 16,* 314-326.

Boccaccini, M. T., Kan, L. Y., Rufino, K. A., Noland, R. M., Young-Lundquist, B. A., & Canales, E. (2016). Correspondence between correctional staff and offender ratings of adaptive behavior. *Psychological Assessment*, *28*, 1608-1615. doi: 10.1037/pas0000333

Brewer, W. F. (1986). What is autobiographical memory? In D. C. Rubin (Ed.). *Autobiographical memory* (p. 25-49). New York: Cambridge University Press.

Bruininks, R., Woodcock, R.W., Weatherman, R.F. & Hill, B.K. (1984). *Scales of independent behavior: Instructor's edition.* Allen, TX: DLM Teaching resources

Bruininks, R., Woodcock, R.W., Weatherman, R.F. & Hill, B.K. (1996). *SIB-R: Scales of independent behavior—Revised.* Itasca, IL: Riverside Publishing.

Conway, M.A. & Holmes, E. (2008). *Guidelines on memory and the law: Recommendations from the scientific study of human memory*. The British Psychological Society Press.

Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

De Los Reyes, A. (2011). Introduction to the Special Section: More Than Measurement Error: Discovering Meaning Behind Informant Discrepancies in Clinical Assessments of Children and Adolescents. *Journal of Clinical Child & Adolescent Psychology,* 40, 1–9.

De Los Reyes, A. (2013). Strategic objectives for improving understanding of informant discrepancies in developmental psychopathology research. *Development and Psychopathology, 25,* 669–682. http://dx.doi.org/ 10.1017/S0954579413000096

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin,* 131, 483–509.

De Los Reyes, A., & Prinstein, M. J. (2004). Applying depression–distortion hypotheses to the assessment of peer victimization in adolescents. *Journal of Clinical Child and Adolescent Psychology,* 33, 325–335.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E, & Rabinowitz, J. (2015). The Validity of the Multi-Informant Approach to Assessing Child and Adolescent Mental Health. *Psychological Bulletin,*141, 858–900.

Doane, B. M., & Salekin, K. L. (2009). Susceptibility of current adaptive behavior measures to feigned deficits. *Law and Human Behavior*, *33*, 329-343. doi: 10.1007/s10979-008-9157-5

Doll, E. A. (1953). *The measurement of social competence: a manual for the Vineland social maturity scale*. Minneapolis: Educational Test Bureau, Educational Publishers.

Doss v. State, 19 So.3d 690, 713 (Miss. 2009).

Ex parte Briseno, 135 S.W.3d 1 (Tex. Crim. App. 2004).

Federal Rules of Evidence, Pub. Law (1975). 93–595.

Frye v. United States, 293 F. 1013 (D.C. Cir. 1923).

General Electric Co. v. Joiner, 522 U.S. 136 (1997).

Gonzales, N. A., Cauce, A. M., & Mason, C. A. (1996). *Inter-observer agreement in the assessment of parental behavior and parent–adolescent conflict: African American mothers, daughters, and independent observers*. Child Development, 67, 1483–1498.

Goodman, K. L., De Los Reyes, A., & Bradshaw, C. P. (2010). Understanding and using informants' reporting discrepancies of youth victimization: A conceptual model and recommendations for research. *Clinical Child and Family Psychology Review,* 13, 366–383.

Greenspan, S. & Switzky, H.N. (2006). Lessons from the Atkins decision for the next AAMR manual. In H.N. Switzky & S. Greenspan (Eds.), *What is mental retardation? Ideas for an evolving disability in the 21st century.* (pp. 281-300). Washington, D.C.: American Association on Mental Retardation.

Hall v. Florida, 134 S. Ct. 1986 (2014).

Harrison, P.L., & Oakland, T. (2003), *Adaptive Behavior Assessment System – Second Edition,* San Antonio, TX: The Psychological Corporation.

Harrison, P.L., & Oakland, T. (2015), *Adaptive Behavior Assessment System –Third Edition* [Manual]. Torrance, CA: Western Psychological Services.

Hundert, J., Morrison, L., Mahoney, W., Mundy, F., & Vernon,M. L. (1997). Parent and teacher assessments of the developmental status of children with severe, mild/moderate, or no disabilities. *Topics in Early Childhood Special Education,* 17, 419-434.

Jobe, J. B., White, A. A., Kelley, C. L., Mingay, D. L., Sanchex, M. J., & Loftus, E. F. (1990). Recall strategies and memory for health care visits. *Milbank Quarterly, 68,* 171-189.

Kessler, R. C., Wittchen, H.U., Abelson, J., & Zhao, S. (2000). Methodological issues in assessing psychiatric disorders with self-reports. In A. Stone, J. Turrkan, C. Bachrach , J. Jobe, H. Kurtzman, & V. Cain (Eds.), *The Science of Self-Report: Implications for Research and Practice* (p. 229-255). Mahwah, NJ: Erlbaum Associates.

Lampinen, J. M., Copeland, S. M., & Neuschatz, J. S. (2001). Recollections of things schematic: Room schemas. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 5, 1211-1222. DOI: 10.1037//0278-7393.27.5.1211.

Loftus, E. (2003). Our changeable memories: Legal and practical implications. *Nature Reviews Neuroscience*, *4*, 231-234.

Kraemer, H. C.,Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce,W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry*, 160, 1566–1577.

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the emergence of conceptual knowledge during human decision making. *Neuron, 63,* 889–901.

Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999).

Lapouse, R., & Monk, M. A. (1958). An epidemiologic study of behavior characteristics of children. *American Journal of Public Health*, 48, 1134–1144.

Macvaugh, G., & Cunningham, M. D. (2009). Atkins v. Virginia: Implications and recommendations for forensic practice. *Journal of Psychiatry and Law*, 37, 131-187.

Markus, H. & Zajonc, R. B. (1985). The cognitive perspective in social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (3rd ed., P. 137-230). New York: Random House.

McDonald, C. A., Lopata, C., Donnelly, J. P., Thomeer, M. L., Rodgers, J. D., & Jordan, A. K. (2016). Informant Discrepancies in Externalizing and Internalizing Symptoms and

Adaptive Skills of High-Functioning Children with Autism Spectrum Disorder. *School Psychology Quarterly*, 31, 467–477.

Means, B., & Loftus, E.E. (1991). When personal history repeats itself: Decomposing memories for recurring events. *Applied Cognitive Psychology, 5,* 297-318.

Menon. A. (1994). Judgments of behavioral frequencies: Memory search and retrieval strategies. In N. Schwartz & S. Sudman (Eds.), *Autobiographical memory and the validity of retrospective reports* (p. 161-172). New York: Springer-Verlag.

Merlino, M.L., Springer , V. Seaman Kell, J., Hammond, D., Sahota, E. & Haines, L. Meeting The Challenges Of The *Daubert* Trilogy: Refining And Redefining The Reliability Of Forensic Evidence. Tulsa Law Review, Volume 43, Issue 2 *Daubert,(2007).*

Moore v. Texas, 136 S.Ct. 2407 (2016).

Paavola, E. C. (2016, April). *Atkins Update: The Current State of Atkins Litigation.* Paper presented at 13TH National Seminar on the Development and Integration of Mitigation Evidence. New Orleans, LA.

Paclawskyj, T. R., Matson, J. L., Rush, K. S., Smalls, Y., & Vollmer, T. R. (2000). Questions about Behavioral Function (QABF): A behavioral checklist for functional assessment of aberrant behavior. *Research in Developmental Disabilities*, 21, 223–229. doi:10.1016/S0891-4222(00)00036-6

Preston, A., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology, 23,* R764–R773.

Renk, K. (2005). Cross-informant ratings of the behavior of children and adolescents: The "gold standard." Journal of Child and Family Studies, 14, 457–468.

Roper v. Simmons, 543 U.S. 551 (2005).

Schacter, D. L. (2012). Constructive memory: Past and future. *Dialogues on Clinical Neuroscience,* 14, 7-18.

Schalock, R. L., Borthwick-Duffy, S. A., Bradley, V. J., Buntinx, W. H. E., Coulter, D. L., Craig, E. M., Gomez, S. C., Lachapelle, Y., Luckasson, R., Reeve, A., Shogren, K. A., Snell, M. E., Spreat, S., Tasse, M. J., Thompson, J. R., Verdugo-Alonso, M. A., Wehmeyer, M. L., & Yeager, M. H. (2010). *Users guide to intellectual disability: Definition, classification, and systems of supports.* Washington, DC: American Association on Intellectual and Developmental Disabilities.

Shea, N., Krug, K., & Tobler, P. N. (2008). Conceptual representations in goal-directed decision making. *Cognitive, Affective & Behavioral Neuroscience, 8(4),* 418–428.

Shogren, K. A., & Rojahn, J. (2003). Convergent reliability and validity of the Questions about Behavioral Functions and the Motivation Assessment Scale: A replication study. *Journal of Developmental and Physical Disabilities*, 15, 367–375. doi:10.1023/A:1026314316977

Sigafoos, J., Kerr, M., & Roberts, D. (1994). Interrater reliability of the Motivation Assessment Scale: Failure to replicate with aggressive behavior. Research in Developmental Disabilities, 15, 333–342. doi: 10.1016/0891-4222(94)90020-5

Sparrow, S.S., Cicchetti, D.V., & Balla, D.A. (2005). *Vineland Adaptive Behavior Scales-Second Edition*, San Antonio, TX: The Psychological Corporation.

Sparrow, S.S., Cicchetti, D.V., & Saulnier, C. A. (2016). *Vineland Adaptive Behavior Scales-Third Edition*, San Antonio, TX: The Psychological Corporation.

Stevens, S. S. (1946). On the Theory of Scales of Measurement, *Science*, 103, 677-680.

Stevens, S. S. (1960). Philosophical and Foundational Issues in Measurement. C. Wade Savage, & Philip Ehrlich (Eds), New Jersey: Lawrence Erlbaum

Szatmari, P, Archer, L., Fisman, S., & Streiner, D. L. (1994). Parent and Teacher Agreement in the Assessment of Pervasive Developmental Disorders. *Journal of Autism and Developmental Disorders*, 24, 703-717.

Taber, S. M. (2010). The veridicality of children's reports of parenting: A review of factors contributing to parent–child discrepancies. Clinical Psychology Review, 30, 999–1010.

Thomas v. Allen 614 F. Supp. 2d 1257 (2009).

U.S. v. Hardy, 762 F.Supp.2d 849, 854 (E.D. La. 2010).

U.S. v. Smith, 790 F. Supp.2d 482 (E.D. La. 2011).

Van Kesteren, M. T. R., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., & Fernández, G. (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: From congruent to incongruent. *Neuropsychologia, 1–8.*

Voelker, S., Shore, D., Hakim-Larson,, J., & Bruner, D. (1997). Discrepancies in parent and teacher ratings of adaptive behavior of children with multiple disabilities. *Mental Retardation,* 35, 10-17.

Wechsler, D. (2008). *Wechsler Intelligence Scale for Children – Fourth Edition*, San Antonio, TX: The Psychological Corporation.

Widaman, K. F., & Siperstein, G. N. (2009). Assessing adaptive behavior of criminal defendants in capital cases: A reconsideration. *American Journal of Forensic Psychology*, 27, 15-32.

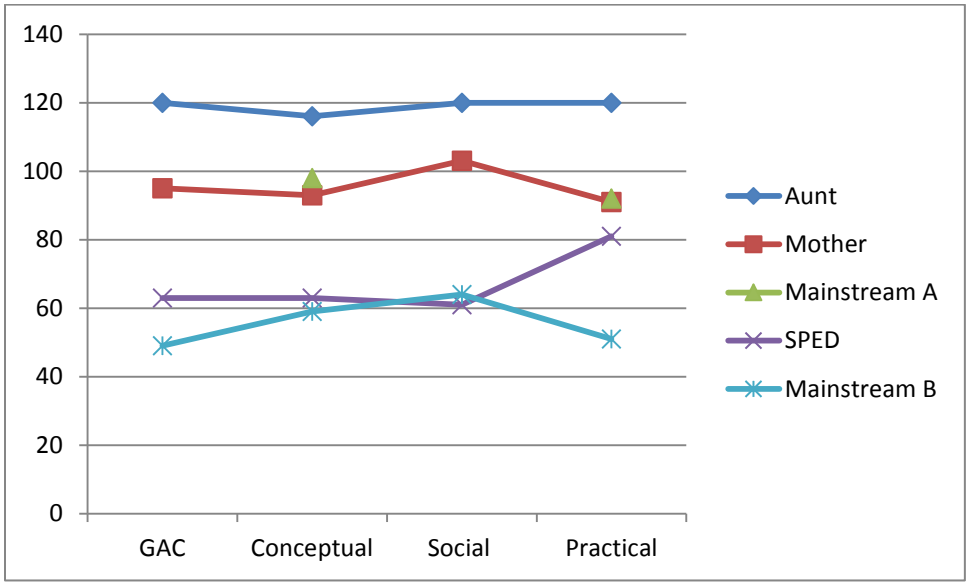
Wiley v. *Epps*, 625 F.3d 199 (5th Cir. 2010).

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside Publishing.

Young, B., Boccaccini, M.T., Conroy, M.A., & Lawson, K. (2007). Four practical and conceptual assessment issues that evaluators should address in capital case mental retardation evaluations. *Professional Psychology: Research and Practice, 38*, 169-178. doi: 10.1037/0735-7028.38.2.169

Zarcone, J. R., Rodgers, T. A., Iwata, B. A., Rourke, D. A., & Dorsey, M. F. (1991). Reliability analysis of the Motivation Assessment Scale: A failure to replicate. Research in Developmental Disabilities, 12, 349–360. doi:10.1016/0891-4222(91)90031-M

Figure 1. *ABAS-II Composite Scores from Multiple Raters from Case 1.*



Note: SPED = Special Education Teacher. GAC = General Adaptive Composite.

Table 1. *Composite Scores and Descriptive Category from Multiple Raters from Case 1.*

Standard Scores by Rater

	GAC	Conceptual	Social	Practical
Aunt	120	116	120	120
Mother	95	93	103	91
Mainstream Teacher A	**	98	**	92
Special Education Teacher	63	63	61	81
Mainstream Teacher B	49	59	64	51

Descriptive Classifications of Skill by Rater

	GAC	Conceptual	Social	Practical
Aunt	Superior	Above Average	Superior	Superior
Mother	Average	Average	Average	Average
Mainstream Teacher A	**	Average	**	Average
Special Education Teacher	Extremely Low	Extremely Low	Extremely Low	Below Average
Mainstream Teacher B	Extremely Low	Extremely Low	Extremely Low	Extremely Low

Note: GAC = General Adaptive Composite. **Indicates missing data, as too many items were "guessed."

Figure 2. *SIB-R Standard Scores from Multiple Raters for Case 2.*

Table 2. *Standard Scores, Age Equivalents, and Descriptive Classifications for the SIB-R from Multiple Raters for Case 2.*

Standard Scores by Rater

	Broad Independence	**Motor Skills**	**Social Interaction and Communication Skills**	**Personal Living Skills**	**Community Living Skills**
Mother	48	70	58	57	45
Aunt	60	83	47	89	57
Sister	85	78	64	113	93

Age Equivalents by Rater

	Broad Independence	**Motor Skills**	**Social Interaction and Communication Skills**	**Personal Living Skills**	**Community Living Skills**
Mother	8-4	9-7	7-6	8-6	7-9
Aunt	9-9	12-0	6-1	12-5	8-10
Sister	13-5	11-0	8-8	22-0	14-11

Descriptive Classifications of Skill by Rater

	Broad Independence	**Motor Skills**	**Social Interaction and Communication Skills**	**Personal Living Skills**	**Community Living Skills**
Mother	Limited	Limited	Limited to Very Limited	Limited to Very Limited	Limited to Very Limited
Aunt	Limited	Limited to Age-appropriate	Very Limited	Limited to Age-appropriate	Limited
Sister	Limited to Age-appropriate	Limited to Age-appropriate	Limited	Age-appropriate to Advanced	Limited to Age-appropriate

Figure 3. *SIB-R Composite Scores from Multiple Raters from Case 3.*

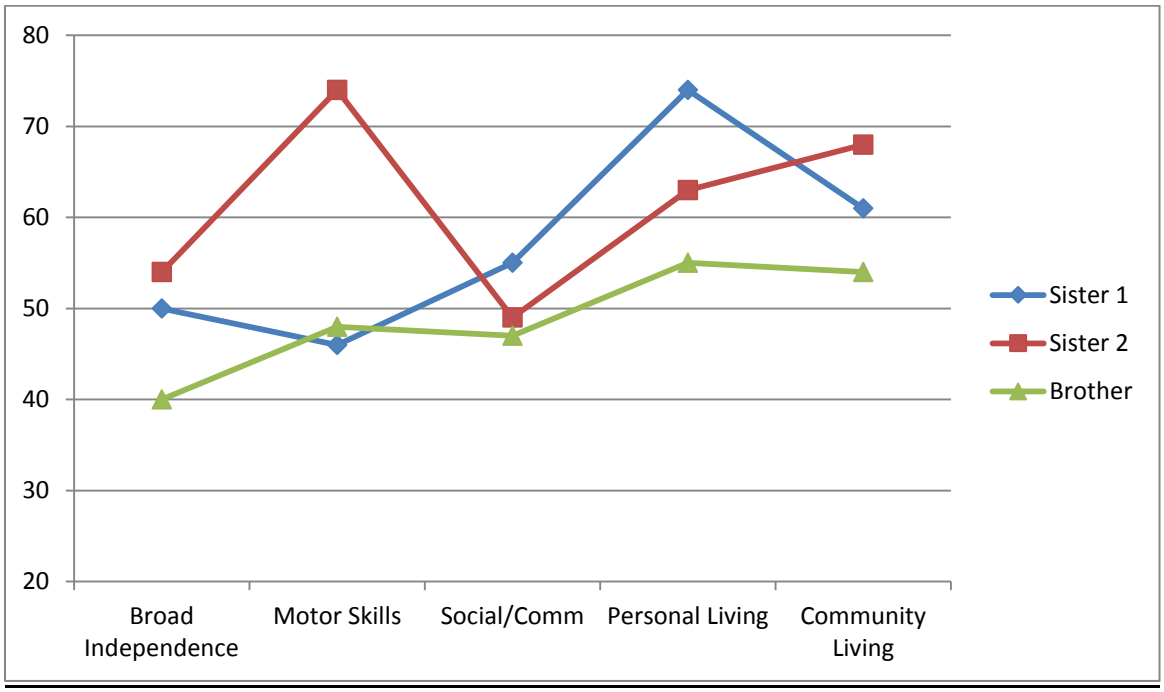


Table 3. *Standard Scores, Age Equivalents, and Descriptive Classifications from Multiple Raters from Case 3.*

Standard Scores by Rater

	Broad Independence	**Motor Skills**	**Social Interaction and Communication Skills**	**Personal Living Skills**	**Community Living Skills**
Sister 1	50	46	55	74	61
Sister 2	54	74	49	63	68
Brother	40	48	47	55	54

Age Equivalents by Rater

	Broad Independence	**Motor Skills**	**Social Interaction and Communication Skills**	**Personal Living Skills**	**Community Living Skills**
Sister 1	8-6	6-8	7-0	11-4	9-6
Sister 2	9-0	10-3	6-3	9-6	10-7
Brother	7-5	6-10	6-1	8-3	8-8

Descriptive Classifications of Skill by Rater

	Broad Independence	**Motor Skills**	**Social Interaction and Communication Skills**	**Personal Living Skills**	**Community Living Skills**
Sister 1	Limited	Limited to very limited	Limited to very limited	Limited	Limited
Sister 2	Limited	Limited	Limited to very limited	Limited	Limited
Brother	Limited to very limited	Limited to very limited	Very limited	Limited to very limited	Limited to very limited