

Psychometrics in the courtroom: A matter of life and death?

***DRAFT document – Please do not cite without permission***

Chad W. Buckendahl  
*Buros Center for Testing*  
*University of Nebraska – Lincoln*

Paper presented at the annual meeting of the National Council on Measurement in  
Education, Chicago, IL

April 12, 2007

## Abstract

The U.S. Supreme Court issued a landmark ruling in *Atkins v. Virginia* (2002) holding that the execution of mentally retarded individuals violated the eighth amendment's prohibition against cruel and unusual punishment. In *Atkins*, the defendant's death sentence was overturned after evidence demonstrated that he had scored a 59 on an intelligence quotient (IQ) test. This ruling has led to additional challenges related to the determination of a defendant's IQ. These challenges also pose educational, legal, psychological, and psychometric questions. This paper discusses some of these issues in the context of a recent case, *Vela v Nebraska* (2006). Expert witnesses in *Vela* testified about the psychometric properties of the tests at issue in this case, intended uses of the tests, administration requirements, and score interpretation. Implications for practitioners in educational settings are also discussed.

Keywords: intelligence testing, legal issues, validity

## Psychometrics in the courtroom: A matter of life or death?

In the measurement community, we are often confronted with questions regarding appropriate uses for test scores, particularly in the context of high stakes decisions and the consequences that may emerge from these uses. The use of test scores in education to make individual decisions about students' proficiency or graduation eligibility may not be as perceived as critical as the decisions made within admissions, licensure, certification, or employment testing; however, these are all areas that could be characterized as high stakes uses.

Intelligence test scores may be used in educational settings to identify students for special programs. For example, students that earn high scores on an intelligence test may be identified for gifted programs; whereas students that score lower may be identified for remedial or special education programs. Although perhaps not intended for high stakes purposes, students that are identified for these programs may receive additional attention or services based on these test scores. For example, a woman in Tacoma, WA recently admitted that she coached her two children from an early age to fake mental retardation to receive Social Security benefits totaling approximately \$280,000 over two decades (Associated Press, 2007). Thus, the incentives for performing poorly or well on an intelligence test may be great.

A recent court decision has raised the stakes for the use of intelligence test scores in some criminal proceedings. The U.S. Supreme Court issued a landmark ruling in *Atkins v. Virginia* (2002) holding that the execution of mentally retarded individuals violated the eighth amendment's prohibition against cruel and unusual punishment. In this case, the defendant's death sentence was overturned after evidence demonstrated that he had scored a 59 on an intelligence quotient (IQ) test suggesting that he was mentally retarded. This recent ruling has led to additional challenges that focus on the determination of a defendant's IQ and its influence on sentencing decisions. These challenges pose educational, legal, psychological, and psychometric questions.

This paper focuses primarily on the results of intelligence tests and discusses some of these issues in the context of a recent case, *Vela v Nebraska* (2006). Expert witnesses in *Vela* testified about the psychometric properties of the tests at issue in this case, intended uses of the tests, administration requirements, and score interpretation. Additional research directions and implications for practitioners are also discussed.

### *Background of the Case*

Erick Vela and three other defendants were convicted of killing five people during a 2002 robbery at a U.S. Bank branch in Norfolk, Nebraska. After juries found three of the defendants guilty and eligible for the death penalty, attorneys for Vela filed a motion to assert that their client was mentally retarded and therefore ineligible for the death penalty. The motion was based on the defendant's interpretation of Nebraska's statute and the results of multiple intelligence tests that were administered to the defendant following his arrest and conviction.

Experts retained by the defense (two forensic psychologists and one school psychologist) administered three intelligence tests to Vela on three separate occasions

over a one year period. The primary intelligence tests that were used were the Wechsler Abbreviated Scale of Intelligence (WASI, 1999), the Wechsler Adult Intelligence Scale – Third Edition (WAIS-III, 1997), and the Stanford-Binet Intelligence Scale – Fifth Edition (SB5, 2003). At each administration, a different administrator and a different test or set of tests was used. More important, different scores were observed that could result in different conclusions about the defendant’s intelligence. Specifically, it was only after the third set of tests was administered did Vela’s test scores suggest that he may be mentally retarded using the state’s statutory language. Therefore, the court’s interpretation of the validity of these test scores was critical to the judge’s ruling.

During a hearing to determine whether the defendant was mentally retarded, both the defense and prosecution utilized expert witnesses to assist in the interpretation of these scores. The defense’s primary expert witness was the school psychologist that had administered the third set of intelligence tests to Vela. Conversely, the prosecution’s expert witnesses included both of the defense’s forensic psychologists, two of their own forensic psychologists, and a psychometrician. At the hearing, these witnesses were called to testify about the psychometric properties of the tests at issue in this case, intended uses of the tests, administration requirements, and score interpretation.

### *Selected Literature*

The interpretation of the results of the intelligence tests in this capital punishment case was based on clinical, legal, and psychometric characteristics that the judge considered. Therefore, this section briefly discusses literature related to each of these perspectives. We first discuss the definition and measurement of mental retardation as promulgated by the clinical psychology community. Then, we describe the legal literature that was important to this case. Finally, the interpretation of validity in the context of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) is also included.

The clinical interpretation of mental retardation is defined in the *Diagnostic and Statistical Manual of Mental Disorders – 4<sup>th</sup> Edition* (DSM-IV; American Psychiatric Association, 1994) as having three elements. These elements are: 1) the individual demonstrates sub-average intellectual functioning; 2) the individual demonstrates significant limitations in adaptive functioning; and 3) these observations are manifest before age 18. The first two elements have been further clarified for clinicians.

Sub-average intelligence is generally defined as an intelligence quotient (IQ) of approximately 70 or lower. This value represents a scale score that is typically two standard deviations lower than the mean score on commonly used commercially-available intelligence tests. The uncertainty in a clear cut score on an intelligence measure in the clinical definition represents an understanding of measurement error that is present in the estimation of an individual’s scores. Because the standard error of measurement associated with the full scale score is approximately 2.5 for many of these tests, the expected range of scores around 70 ranges from 65 to 75 using a 95% confidence interval. Because there are multiple elements in the determination of mental retardation, consideration of the measurement error allows the clinical decision to be compensatory.

Adaptive functioning is often defined as the ability to communicate, care for oneself, respond, and react to activities in daily life. Evaluating an individual's adaptive functioning may occur through observations, interviews with an individual or people who know the individual, or administration of standardized instruments designed to measure these characteristics (e.g., Vineland Adaptive Behavior Scales). The results of standardized instruments are typically reported as scale scores and considered in the judgment about whether an individual is mentally retarded.

From a clinical perspective, it is important to understand that there is not a fixed cut score on intelligence or adaptive functioning that determines mental retardation. However, the judgment is able to consider both sets of data allowing an individual to compensate for intelligence that might range from 70 to 75, but is then offset by lower performance on adaptive functioning measures. The alternate scenario is also possible. In both instances, the interpretation of the scale scores is in the context of relative performance to a norm sample that defines an expected distribution of scores in the population.

The legal definition of mental retardation often follows the clinical definition in many states. Nebraska has statutory language that allows for a presumption of mental retardation when a reliably administered intelligence test yields a 70 or lower. For the use of intelligence test scores in this context there is also caselaw that is relevant to this discussion.

In *Bowden v. Georgia* (1982) the defendant brought evidence from previous psychological testing about his intellectual functioning in an appeal of a death sentence. In this case, the defendant brought a psychologist's report from 1966 that found that he was functioning within the lower limits of mild retardation. Although the defendant had a history of low ability in academic and social settings, the court in *Bowden* rejected the argument that this prior information mitigated the sentence. The rationale for the rejection was that the evidence did not suggest that the defendant was insane when the crime was committed or that he was legally incompetent at trial.

It was not until *Atkins v. Virginia* (2002) that the Supreme Court ruled that a defendant who scored a 59 on an intelligence test should not be eligible for the death penalty as it would violate the Eighth Amendment's prohibition against cruel and unusual punishment. The Court did not rely solely on the defendant's performance on the intelligence test as the sole determinant for their decision; they also considered the additional elements of the clinical definition of mental retardation (e.g., adaptive functioning, age of onset) to support their opinion.

Buckendahl and Hunt (2005) noted that there are sometimes different expectations when comparing legal versus professional standards. When statutes or caselaw are not available to guide the judgments, the courts will often defer to the respective professional community for guidance on what is appropriate best practice. The interpretation of evidence may also be variable when comparing the professional community to the legal standard. For example, most standard setting literature in educational or credentialing contexts recommend criterion-referenced decisions regarding cut scores, rather than normative cut scores that predetermine a certain proportion of examinees below the cut score. However, in employment testing, norm-referenced cut scores may be appropriate depending on the use and interpretation of the scores.

In a review of some prominent court cases that focused on the interpretation and use of test scores, Sireci and Parker (2006) illustrated how the legal opinions rendered in these cases were responsive to expectations in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). In many of these cases, the judges relied on information from the technical documentation and our professional standards as guiding principles for their decisions. This represents a strong statement for high stakes testing programs that are concerned about legal defensibility.

Of particular relevance to this case is the consideration of the characteristics and interpretation of intelligence test scores. As mentioned above, intelligence tests are developed using norm reference groups to establish an expected score distribution to which observed scores are then compared. Flynn (1987) examined longitudinal changes in intelligence scores and concluded that improved social conditions, specifically in developed countries, have contributed to a rise in the level of intelligence in the general population. These findings have implications for the norm samples and the subsequent interpretation of these scores. Often characterized as the ‘Flynn effect’, the interpretation of scores may not be stable because the construct in the general population may not be invariant. Thus, interpreting an examinee’s scores using an outdated set of norms may overestimate his or her ability relative the current population, if the level of intelligence continues to increase at a known, systematic rate.

Wicherts, et al. (2004) also examined the characteristics of intelligence scores over time. In one of the studies they conducted, they evaluated measurement invariance between cohorts from 1944 and 1984. The results from this study suggested that the underlying factor structure was different for these two groups. As the authors noted, additional work in the area is needed, but it may add another layer to discussions about explaining reasons for observations that are attributed to the Flynn effect.

From our review of relevant literature, it is apparent that the courts have relied on guidance from professional communities, particularly in clinical psychology, in defining mental retardation. Recent statutes and caselaw have generally drawn heavily on the clinical definition of mental retardation to support their decisions. Additional factors such as using norm-referenced cut scores and possible definition (See Sternberg, 1996) and variability of the intelligence construct in the general population remains an issue that the courts have not yet fully considered in their use of test scores for a new high stakes purpose. In the next section we describe the tests that were administered to Vela as an illustration of how intelligence tests are currently being used in legal proceedings.

### *Methods*

A series of tests were administered to Vela during a one year timeframe. The first set of tests was administered in July 2003, the second set of tests was administered in November 2003, and the third set of tests was administered in July 2004. Table 1 shows the time sequence of test administration with selected tests that were administered.

Table 1. Selected tests administered to the defendant by date

---

<u>Date</u>	<u>Intelligence</u>	<u>Adaptive Behavior</u>	<u>Malingering</u>
-------------	---------------------	--------------------------	--------------------

July 2003	WASI	none	none
November 2003	WAIS-III	none	TOMM, VSVT, VIP, 21-item
July 2004	SB5	Vineland (2004-05)	none

Intelligence Tests

The three intelligence tests that were administered to Vela are commonly used with adult populations. Each of the expert witnesses selected by the defense received information about the previous test administration(s) and results prior to their respective administration. These defense experts (two forensic psychologists and one school psychologist) were experienced with selecting, administering, scoring, and interpreting the intelligence tests used in this case. A brief description of the intended population, use, and some relevant cautions of each intelligence test is provided here.

The Wechsler Abbreviated Scale of Intelligence (WASI, 1999) is described by the publisher as a “short, reliability measure of intelligence.” The intended population of the instrument is for children and adults ages 6 to 89. The technical manual does note a caution about interpreting the scores. Specifically, the scores are not intended to make diagnostic or placement decisions; or replace more comprehensive measures of intelligence.

The Wechsler Adult Intelligence Scale – Third Edition (WAIS-III, 1997) is described by the publishers as designed “to assess the intellectual ability of adults.” The intended population for the instrument is adults ages 16 to 89. The WAIS-III technical manual cautions against using only the scores from the test to diagnose or preclude low intellectual abilities. An additional caution from the WAIS-III technical manual is the need for the test administrator to consider the examinee’s effort or motivation in their interpretation of the scores. In this case, the defendant may not be motivated to perform well if he is aware of the potential consequences of performing well on the test.

The third intelligence test administered was the Stanford-Binet – Fifth Edition (SB5, 2003) which was designed by the publisher to assess “intelligence and cognitive abilities.” The intended population for the instrument is children and adults with ages ranging from approximately 2 to 89. The SB5 notes a caution in their technical manual that was relevant to this case. Specifically, that “when the context of the assessment and the examinee’s background is influenced by such factors as communication disorders, learning disabilities, autism, or non-English background, the Non-verbal IQ score may be a better indicator of global cognitive potential.” This caution became more relevant when examining the scores from the intelligence tests.

On these three intelligence tests, scores are reported on a scale that has a mean of 100 and a standard deviation of 15. Table 2 shows the verbal, performance (non-verbal), and composite scores obtained by Vela for each of the intelligence tests over the three administration periods.

Table 2. Observed scores for each intelligence test across administrations.

<u>Date</u>	<u>Test</u>	<u>Verbal</u>	<u>Performance</u>	<u>Full Scale</u>
July 2003	WASI	82	94	87
Nov. 2003	WAIS-III	75	78	75
July 2004	SB5	56	79	66

#### Adaptive Behavior Test

Because the professional psychology community does not recommend diagnosing mental retardation using an intelligence test alone, an adaptive behavior test was also administered by one of the defense experts and also by one of the prosecution experts. The Vineland Adaptive Behavior Scales (1985) were administered as part of the evaluation by the third defense expert in 2004 and by the prosecution's expert in 2005. The defense expert used Vela's sister as a subject, whereas the prosecution's expert used two of Vela's friends as subjects to triangulate the results of the test.

#### Malingering Tests

In addition to the intelligence and the adaptive behavior tests, one of the defense's forensic psychologists also administered some malingering tests to ascertain whether the defendant was putting forth his best effort. In instances where there might be an incentive for putting forth less than one's best effort (e.g., worker's compensation, receiving special services), the use of these tests may be an important component to the validity evidence. The results of these tests were also used to mitigate the confidence in the intelligence test scores and also to inform the conclusions of the evaluation. As shown in Table 1 above, no malingering tests were administered during the July 2003 or July 2004 administrations. A series of malingering tests were included in the test battery in November 2003. These tests were the Test of Memory Malingering (TOMM), Victoria Symptom Validity Test (VSVT), Validity Indicator Profile (VIP), and the 21-item Test. Results of these tests during the November 2003 administration were inconclusive as to whether the defendant was motivated to perform on the cognitive tests that were part of the battery.

#### Additional evidence considered

Although the discussions in this hearing focused primarily on the results and appropriate interpretation of the intelligence tests in the context of Nebraska's statutory language, both the prosecution and defense brought additional evidence that might be used to support their respective validity argument (Kane, 1992). These are briefly mentioned here because of the educational relevance of some of the evidence.

Although determined to be less relevant in the decision, the defense's third expert witness also administered other tests of achievement and intelligence. These were the Peabody Picture Vocabulary Test (PPVT), Comprehensive Test of Non-Verbal Intelligence (CTONI), and the achievement subtests of the Woodcock-Johnson III (WJ-III). The defense brought in elementary schoolteachers who had taught Vela when he was

growing up in Inglewood, CA. They were asked to comment on his intellectual functioning and his social skills in the classroom. The defense also had Vela's sister testify as to his adaptive functioning and his ability to take care of himself. Her testimony suggested that he needed substantive assistance in both academic and social settings. The defense also brought in fellow inmates from the prison that had observed Vela and testified that they helped him make phone calls and write letters.

The prosecution countered with educators from the school district who had access to Vela's academic transcripts that noted he was not identified as mentally retarded or for special education programs; however, he had taken remedial coursework in core subject areas. His course grades were also included in this transcript. The prosecution also brought guards from the prison to testify about Vela's ability to write notes (called *kites*) to the guards regarding specific requests (e.g., a haircut, a book, a phone call). These witnesses were asked to describe their observations of his social and behavioral functioning in the prison.

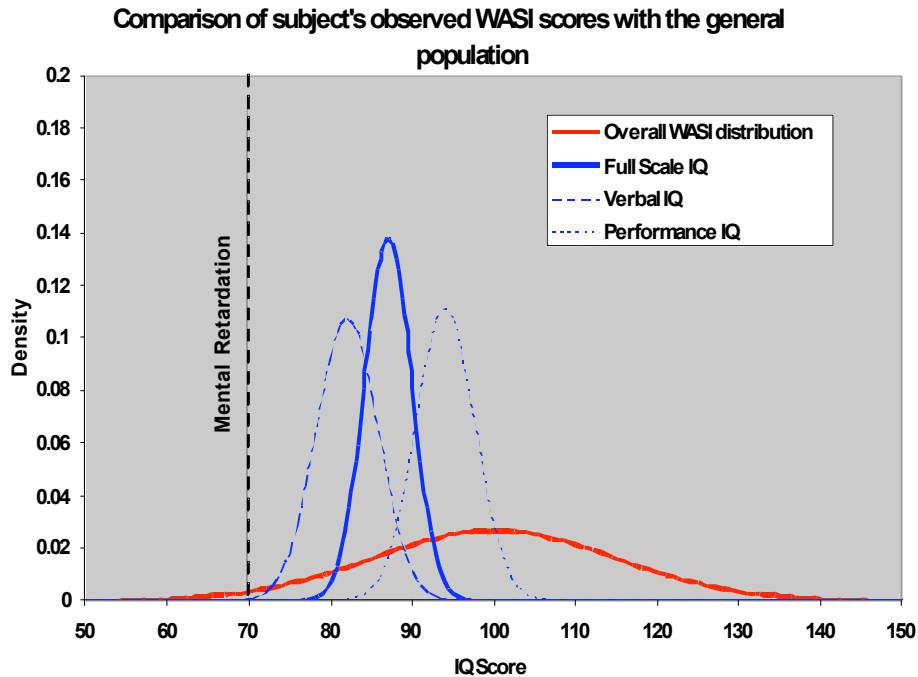
### *Results*

In this section, we discuss the results of the intelligence, adaptive behavior, and malingering tests that were administered and admitted as evidence as part of this case.

At the hearing, experts for the defense and prosecution discussed the results of Vela's test scores and the appropriate interpretations of each in the context of the requirements of the publisher. At the judge's request, specific information was requested regarding how the court might interpret the differences in observed scores across the intelligence tests that were administered in this case. To respond to this request, we prepared graphical representations of Vela's test scores in comparison to a distribution of scores that we might observe in the general population. Also in these displays, we showed where Vela's observed score was located with respect to the statutory cut score (i.e. 70). Then, using the standard errors of measurement reported by the publishers, we calculated the probability that Vela's true score was above (or below) the cut score. We also did this for each of the sub-scales for reasons that are better explained in the discussion.

Figure 1 shows the defendant's observed scores from the Wechsler Abbreviated Scale of Intelligence (WASI, 1999). The defendant's full scale score, 87, is shown in relationship to other reference data. First, the full scale score, with its associated estimated standard error is compared with the full scale score distribution with a mean of 100 and a standard deviation of 15. Second, the full scale score is compared with the two observed subscale scores (Verbal – 82 and Performance – 94). Finally, the full scale score is compared with the statutory cut score (70) to illustrate the relationship of the defendant's expected score distribution to the decision point.

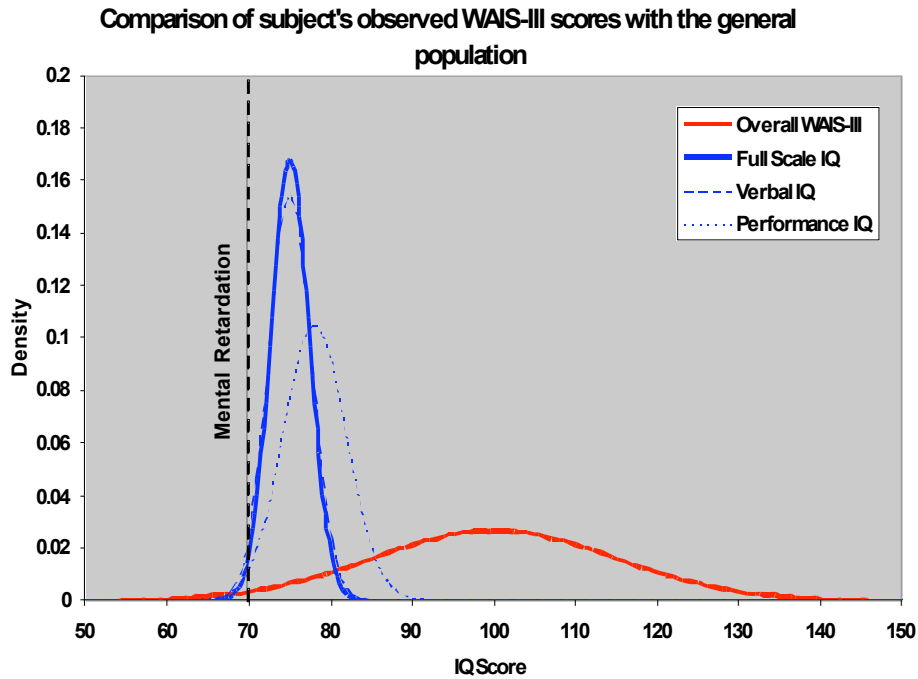
Figure 1.



In Figure 1, we see that the observed full scale score distribution is almost entirely above the decision point and within one standard deviation of the mean of the full population distribution. In calculating the probability that the defendant's true score was below the decision point given the observed scores, we noted that this produced a small value (approximately 1 in 500 million). However, as noted in the technical manual, the use of this test was not intended for diagnosis of mental retardation or placement into special programs.

Figure 2 illustrates the defendant's observed scores from the Wechsler Adult Intelligence Scale – Third Edition (WAIS-III, 1997). The defendant's full scale score, 75, is shown in relationship to other reference data. First, the full scale score, with its associated estimated standard error is compared with the full scale score distribution with a mean of 100 and a standard deviation of 15. Second, the full scale score is compared with the two observed subscale scores (Verbal – 75 and Performance – 78). Finally, the full scale score is compared with the statutory cut score (70) to illustrate the relationship of the defendant's expected score distribution to the decision point.

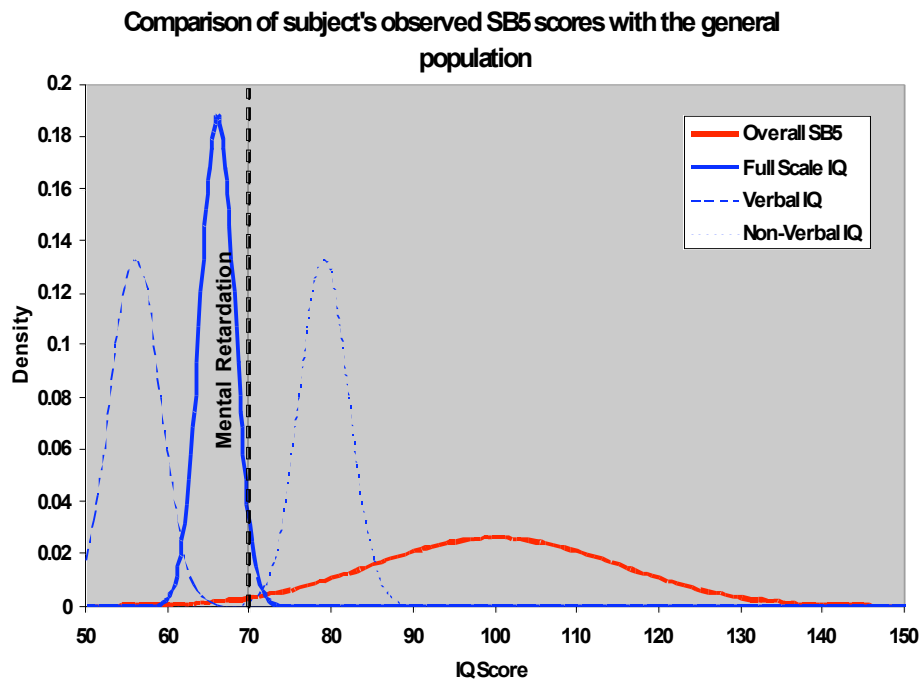
Figure 2.



In Figure 2, we see that the observed full scale score distribution is closer to the decision point than the scores from the first intelligence test that was administered. The probability that the defendant's true score was below the decision point given the observed scores was now not as remote. Specifically, this was calculated as 1.7%. The interpretation of the scores was also cautioned in this test's technical manual as encouraging the examiner to consider motivation and other factors in the interpretation of the scores. The full scale score performance on this test did not meet the statutory language for a presumption of mental retardation; however, it was within the range suggested by some professional psychology literature when considered with other evidence.

Figure 3 illustrates the defendant's observed scores from the Stanford-Binet Intelligence Test – Fifth Edition (SB5, 2003). The defendant's full scale score, 66, is shown in relationship to other reference data. First, the full scale score, with its associated estimated standard error is compared with the full scale score distribution with a mean of 100 and a standard deviation of 15. Second, the full scale score is compared with the two observed subscale scores (Verbal – 56 and Performance – 79). Finally, the full scale score is compared with the statutory cut score (70) to illustrate the relationship of the defendant's expected score distribution to the decision point.

Figure 3.



In Figure 3, we see that the observed full scale score distribution was almost completely below the decision point when compared to scores from the first intelligence tests that were administered. The probability that the defendant's true score was below the decision point given the observed scores was now quite likely. Specifically, this was calculated as 96.9%.

Two cautions from the manual are worth noting here. First, the difference between the two subscale scores raises a flag. The 23 point difference in the two subscales was only observed in 1.9% of the norm sample. When significant differences between subscales are observed, the Stanford-Binet technical manual encourages examiners to be cautious about interpreting a full scale IQ score as the best representation of the examinee's ability. Second, the manual also notes that if the examinee does not have an English speaking background, the non-verbal (Performance) IQ score may be a better indicator of ability. This literacy component was relevant for this defendant.

### *Discussion*

In this section of the paper, the strategies employed by defense counsel in this case, the judge's ruling in *Vela*, implications for practitioners, and future research opportunities in the measurement, legal, and psychological communities that emerged from this case.

### Defense strategy

Based on the results of the three intelligence tests that were the primary focus of the eventual ruling, defense counsel's primary strategy was to highlight Nebraska's

statutory language that presumes mental retardation when a reliably administered intelligence test yields a score of 70 or lower. Because the defendant's results on the third intelligence test administered (Stanford-Binet 5<sup>th</sup> Edition) produced a full scale score of 66, it met the statutory requirement assuming it was reliably administered. The defense would need to rely on a strict interpretation of the statute to support their claim. However, in anticipation that the judge's decision would not rest on one source of evidence, the defense also attempted to make their case using psychometric theory.

Another part of the defense's strategy was to suggest that validity (accuracy) could be obtained without reliability (precision). For the defense, this was a critical attempt to challenge prevailing psychometric theory given the variability in observed scores that the defendant exhibited on the three intelligence tests that were administered. Because there was a lack of convergence in the scores, the defense needed to offer an explanation to the judge for why this occurred. Suggesting that valid score interpretations were possible with limited evidence of reliability was an initial attempt to explain the observed scores.

The second defense approach to explaining the variation in observed scores was a discussion of standard errors each of the scores. This strategy was an attempt to suggest that the scores were not really different. For each of the full scale scores, the standard error of measurement was approximately 2.5 scale score points. Using this value, one of the defense attorneys drew normal curves on flip chart paper and attempted to "link" the defendant's score distributions between the full scale scores observed on the WAIS-III (75) and the SB5 (66). In doing this, the defense was attempting to demonstrate how it was statistically possible that upper end of the SB5 range at +2 SEM (i.e. 71) overlapped with the lower end of the WAIS-III range at -2 SEM (i.e. 70). The conclusion from this analysis was that the defendant's "true" score was approximately 70, even though statistically, the combined probability of these two independent events was small.

Note that from the outset, the defense sought to severely discount (or perhaps ignore) the results of the WASI (87); relying on the manual's caution to not diagnose mental retardation with these scores. However, the defense expert who administered the WASI acknowledged under cross-examination that the likelihood of mental retardation could be screened out by using the scores.

The third defense strategy to explain the observed score differences was to introduce the Flynn (1987) effect that suggested that the differences in scores were due to changes in intelligence in the underlying norm sample. As described above, the Flynn effect hypothesizes that societal and population changes across generations have changed the interpretation of scores on intelligence tests. To account for this theory in the defendant's scores, the defense attributed an approximately 1-point per year reduction in the score when shifting from the 1997 norms used for the WAIS-III versus the 2003 norms used for the SB5. Using the defense's logic, by 2003, the defendant would have exhibited a full scale score of 68 on the WAIS-III which would be quite similar to the score observed on the SB5.

This was an interesting attempt because it spoke to the compliance of the examiner with the test publishers' scoring and interpretation manuals. Although the technical manual acknowledges the potential for the Flynn effect, the scoring and interpretation expectations do not instruct examiners to adjust scores. This lack of direction in the scoring and interpretation sections of the manual may be because it is

unknown whether the effect is uniform or observable over a short period of time when compared with the longitudinal data that Flynn studied. Directions to adjust scores would both recognize and quantify the impact of the effect. Both of these would be controversial without more conclusive evidence to support the decision.

### Judge's ruling

The decision in this case was based on two different legal analyses by the judge. An initial analysis focused on the statutory requirements and the second analysis considered a definition of mental retardation that extended beyond the statute. Each of these analyses included psychometric evidence to support the decision.

In the first analysis, the judge rejected the defense's claim that it met the statutory requirement for presuming mental retardation. Again, the statutory language required that the defendant score a 70 or lower on a reliably administered intelligence test. In the judge's discussion, it was apparent that reliability was interpreted much broader than how the psychometric community might characterize the concept. In his interpretation of the statute, the intent was to support valid decisions about mental retardation, of which reliability might be one component.

In his rejection of the statutory claim, the judge noted that the intelligence test that produced the full scale score below 70 was the third such test administered by the defense. Had it been the first (and perhaps only) intelligence test administered to the defendant, it is likely that these results would have held greater weight in the judge's decision. As a second part of the rationale, the judge suggested that the probability was remote that the 66 the defendant obtained on the SB5 was an accurate representation of the defendant's ability when considered in the context of the two previously administered intelligence tests.

Two additional reasons related to procedural and consequential (cf. Shepard, 1996) validity were also provided by the judge in his rejection of the statutory claim. Given the incentive for performing poorly on the intelligence tests, the judge noted that there were no malingering tests administered during the third battery of tests, even when professional practice would suggest that these be included. Finally, the judge highlighted the third examiner's failure to follow the publisher's interpretation guidelines when drawing conclusions about the defendant's ability as a concern that called into question the results. Thus, on the statutory claim, the defense failed to make their case. However, because the statutory language did not overlap completely with the professional guidelines, the judge also considered the defense's claim more broadly.

In considering the clinical definition of mental retardation, the judge did find that the defendant met the first component (sub-average intelligence) of the three-part criteria based on the results of the WAIS-III administration. The judge relied on the interpretation of standard error and noted that the clinical standard often considers sub-average intelligence scores that range from 65-75. It was interesting to note that the judge did not discuss the scores from the first intelligence test administered by the defense in his conclusion that the defendant met this component. Because it was the second intelligence test administered and resulted in a marked decline from the first test, the credibility of these scores are also quite suspect. On the second component in the clinical definition, the judge determined that the evidence of limitations in adaptive behavior was

insufficient to meet this component. Interestingly, on the third component, the judge noted that Nebraska's statute did not specify an age. Thus, this implies that the clinical requirement regarding the age of onset (18 years old) was judged in this case to not be relevant in the determination of mental retardation.

This third component may be studied in future cases and challenged in appeals in other cases. In this case, it appears that the judge was attempting to reduce the possibility of appeal on some of the intelligence test components by being fairly inclusive of all possible information that could have supported a mental retardation claim. Additional challenges in this area may focus on the validity of the 2<sup>nd</sup> and 3<sup>rd</sup> intelligence tests and rely more heavily on information that is more consistent with the clinical definition.

### Future research

The *Vela* ruling raises a number of additional research questions for practitioners in the measurement, legal, and psychological communities. Some of these areas may overlap with each other and are described here.

When policymakers want to apply a performance standard to a score scale, we often engage in some type of systematic standard setting process drawing from a variety of methods. For mental retardation, the cut score on the intelligence and adaptive behavior scales do not appear to be based on a criterion definition, but rather an examinee's relative position in the population. In most lower stakes situations, this may not be problematic; however, in cases like the one described in this paper, it would be important to the defendant to find a test that is based on a norm group that is "more able" to increase the chances that he or she would fall in the lower tail of the score distribution. The potential for inconsistencies as observed across the three intelligence tests administered in *Vela* increases, particularly when different conclusions are drawn about the results. If these tests are going to be used for these purposes, it may be necessary to explore standard setting methods that are consistent with published literature.

A second set of questions revolve around the basic validity argument of whether scores from these tests should be used for this purpose. In reviewing the technical manuals for each of these intelligence tests, we were unable to find defined, intended purposes beyond using them as measures of intelligence. Although sub-average intelligence is only one component of the clinical definition of mental retardation, it is one that is perhaps more heavily weighted in the decision as suggested in *Atkins* and in Nebraska's statutory language. The intended uses of achievement data were also relevant in *Vela* as school transcripts and observations by teachers were included as testimony. Although one may not be able to anticipate such uses of these data in the future, the validity of these materials was challenged as teachers testified to the low functioning of the defendant in their classes, but were somewhat contradicted by transcripts that documented grades in regular education classes. Had the transcripts shown the defendant being identified as mentally retarded or placed in special education classes; the testimony from the teachers may have been more compelling.

A third set of additional questions is related to all three disciplines and speaks to how intelligence is defined. There continues to be discourse about the measurement and interpretation of intelligence as being distinct from achievement. The current definition within the clinical and legal communities relies heavily on norm-referenced

interpretations of scores in contributing to the decisions about mental retardation. These communities may need to explore the possibility of defining these characteristics in more criterion-referenced terms. As our population becomes more diverse, the need for instruments to be applicable across groups (e.g., norm samples) becomes more important.

Finally, the nature of the intelligence construct across cultures and over time needs much more investigation. Discussions about the presence or absence of the Flynn effect (Flynn, 1987) persist within the literature leading to some interesting debates about the causes of observed changes in intelligence over generations. The invariance of intelligence measures in both situations has been questioned and the rationale is likely more complex than attributing the differences to variation in the norm samples. If the definition of the construct and the stability of the construct are uncertain, then any decisions that result from the use of scores designed to measure the construct will also be tenuous.

### *Conclusions*

The use of tests ranges from low to high stakes situations. As the stakes of the test use/interpretation increase, our expectations for supporting validity evidence also increase. As much as we in the measurement community are responsible for promoting appropriate test use, we need to be equally vigorous in challenging known inappropriate uses of test scores. The case described in this paper illustrates one instance where intelligence and additional educational measurement information were used to support a legal defense in a criminal setting. Similar cases also emerge in civil hearings. Although the judge in this case noted some of the inappropriate test practices and weighed these in his decision, it is possible that future decisions may misinterpret psychometric characteristics of test scores and perhaps establish legal precedents that will run counter to the *Standards*. The judge's use of published literature as a guide was also encouraging, but also suggests that we must continue to support documented evidence for testing programs.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Associated Press (February 27, 2007). Mother coached children to fake retardation. Retrieved February 27, 2007 from <http://www.cnn.com/2007/US/02/27/washington.faked.retardation.ap/index.html>.
- Atkins v. Virginia* (2002). 536 U.S. 304.
- Bowden v. Georgia* (1982). 296 S.E.2d 576.
- Buckendahl, C. W. & Hunt, R. (2005). Whose rules? The relation between the “rules” and “law” of testing. In R. Phelps (Ed.) *Defending standardized testing* (pp. 147-158). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Diagnostic and Statistical Manual of Mental Disorders* 4<sup>th</sup> edition (1994). Washington, DC: American Psychiatric Association.
- Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, 101, 171-191.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Shepard, L. A. (1996). The centrality of test use and consequences for test validity. *Review of Research in Education*, 19, 405-450.
- Sireci, S. G. & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice*, 25(3), 27-34.
- Stanford-Binet – Fifth Edition Technical Manual (2003). Riverside Publishing.
- Sternberg, R. J. (1996). Myths, countermyths, and truths about intelligence. *Educational Researcher*, 25(2), 11-16.
- Wechsler Adult Intelligence Scale – Third Edition Technical Manual (1997). The Psychological Corporation.
- Wechsler Abbreviated Scale of Intelligence Technical Manual (1999). The Psychological Corporation.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, 32, 509-537.

*Vela v. Nebraska* (2006). Case No. CR02-236. Madison, NE: District Court of Madison County.

Vineland Adaptive Behavior Scales Technical Manual (1985). American Guidance Service, Inc.