Open Access Journal of Forensic Psychology

http://www.forensicpsychologyunbound.ws/ – 2009. 1: 51-67

Ethical Implications of Practice Effects on Mental Retardation Claims in Capital Cases

Christopher L. Ray, Ph.D., ABPP, Circleville, Ohio, United States of America.
Chrisray1969@hotmail.com

Abstract: Using the same intelligence test for multiple examinations of a defendant in a capital case can result in practice effects. Strict numerical IQ cutoffs are used in some states to determine whether or not a person is considered mentally retarded. The practice effects of several intelligence tests are reviewed. Findings are presented regarding practice effects for performance versus verbal items on intelligence tests. The importance of time frames, IQ, frequency of re-evaluation, and age with respect to practice effects are examined. Recommendations are discussed concerning the use of intelligence tests in capital cases.

Keywords: intelligence testing, mental retardation, practice effects, IQ cutoffs, capital cases

Introduction

The United States Supreme Court ruled that executing the mentally retarded is a violation of the Eighth Amendment to the Constitution (*Atkins v. Virginia,* 2002). According to Cunningham and Goldstein (2003), prior to this decision, two states and the federal government had prohibited the execution of those with mental retardation. As a result of the *Atkins v. Virginia* decision, Cunningham and Goldstein note that defendants in capital cases are likely to raise questions regarding sub-average intellectual functioning, requiring objective evaluations of their intellectual capacity. Following conviction, assessments focusing on the defendant's intellectual capacity are now commonly used to assist courts in determining the validity of a defendant's claim of mental retardation. The results of assessments focusing on the defendant's intellectual capacity, then, might literally mean the difference between life and death.

In the wake of the *Atkins v. Virginia* (2002) decision, the issue of how to determine whether or not a defendant is mentally retarded has become critically important to courts. In the recent *Bobby v. Bies* (2009) decision, it was held that even mentally retarded individuals sentenced to death pre-*Atkins* can contest their pleas under the postsentencing precedents set forth in *Atkins*. In *Murphy v. Oklahoma* (2002), the Oklahoma Court of Criminal Appeals announced a post-*Atkins* standard for mental retardation to be used in that state's capital cases. The Oklahoma court's mental-retardation standard indicates that in order to be eligible to be considered mentally retarded, the person has to have an intelligence quotient of 70 or below as reflected by at least one scientifically recognized, scientifically approved, and contemporary intelligence quotient (IQ) test. In the *Murphy v. Oklahoma* case, the Oklahoma court

emphasized the use of a set IQ cutoff, which has also been adopted by several other states such as Kentucky and Tennessee.

A key stipulation in *Murphy v. Oklahoma* (2002) is that defendants claiming mental retardation can be evaluated on *at least one* occasion. A complicating factor in multiple assessments of a defendant's IQ is a phenomenon called practice effects. For a myriad of reasons, including practice effects, it is possible for a defendant who scores at or below 70 at the time of initial testing to score above 70 on the same or similar instrument at the time of a subsequent testing. Such a rise in IQ scores could have an important impact on the defendant's fate, particularly in states such as Oklahoma where a defendant must score below 70 to be considered mentally retarded. Duvall and Morris (2006) point out that although the Arizona statute requires that the IQ determination take into account the margin of error for the test administered, this alone does not address the test–retest issue. Moreover, Duvall and Morris indicate that no state statute mandates that an evaluator retesting a defendant's intelligence communicate with prior evaluators to avoid multiple assessment of the defendant's intelligence using identical instruments. With knowledge of practice effects in hand, prosecutors, as Savage (2007) notes, can produce additional scores and other evidence to make the case that an inmate is smart enough to die.

The APA Ethics Code makes it clear that psychologists who perform assessments must consider all relevant test factors (American Psychological Association, 2002). More specifically, standard 9.06 indicates that, when interpreting assessment results, psychologists take into account the various test factors that might affect psychologists' judgments or reduce the accuracy of their interpretations. Moreover, the Ethics Code directs psychologists to indicate any significant limitations of their interpretations. The current review of the practice-effects literature clarifies how practice effects affect intelligence test scores. Moreover, this review examines the extent to which practice effects have an impact on a number of commonly administered intelligence tests and critical factors that influence practice effects. Recommendations are provided on how to examine the intelligence of defendants in capital cases at the time of retesting in an ethically appropriate manner.

Practice Effects Defined

Dawes and Senior (2001) suggest that the lack of perfect reliability that is characteristic of all psychological tests means that no test score remains unchanged and, consequently, clinicians must distinguish between random test-score variability and systematic change. If a clinician becomes convinced that systematic change has occurred, then the basis of that change needs to be determined. Dawes and Senior indicate that improvement in intelligence test scores may be due to recovery of cognitive functions, statistical regression to the mean, or practice effects.

Kaufman (1994) reports that practice effects refer to gains in scores on cognitive tests that occur when a person is retested on the same instrument, or tested more than once on very similar ones. Thorndike (2005) agrees with Kaufman that memory or practice

effects may affect the consistency or stability of a particular score on a test as a description of an individual. Practice-effects gains are due to the experience of having taken the test previously and do not reflect growth or other improvement on the skills being assessed. Such practice effects, suggests Kaufman, denote an aspect of the test itself, a kind of systematic, built-in error that is associated with the specific skills the test measures. Forensic psychologists and other mental-health professionals who assess the intellectual functioning of defendants suspected of mental retardation in capital cases can choose from a wide variety of testing instruments. It is important for mental-health professionals involved in capital cases to be aware of the practice effects that occur in the intelligence tests that they choose to administer.

Practice Effects for the Wechsler Scales

Kaufman states that, in considering the Wechsler scales, the Vocabulary subset typically produces the smallest test-retest gain, and it is usually the most reliable Wechsler subtest. Kaufman also emphasizes that the expected increase of about 5 to 8 points in global IQ renders any score obtained on a retest as a likely overestimate of the person's true level of functioning—especially if the retest is given within about six months of the original test, or if the person has been administered a Wechsler scale (*any* Wechsler scale) several times in the course of a few years. Frumkin (2006) agrees there may be practice effects if intelligence tests such as the Wechsler scales are readministered in less than six months. Kaufman adds that the average range of gain scores makes it feasible for some individuals to gain as much as 15 IQ points due to practice alone.

Matarazzo, Carmody, and Jacobs (1980) reviewed the test-retest stability for the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955). Matarazzo et al. found in their review of 11 test-retest studies of the WAIS that, regardless of large differences in samples and some long intervals of time, the results were consistent in indicating about 2 IQ points of gain on the Verbal Scale and 7 to 8 points on the Performance Scale. The intervals ranged from 1 week to 13 years, the mean ages ranged from 19 to 70 years, and the samples included groups as diverse as brain-damaged elderly, mentally retarded, chronic epileptics, and college students. Matarazzo et al. suggested that Full-Scale IQ should be corrected on retest for an expected gain of 5 points due to practice effects.

The *Wechsler Adult Intelligence Scale-Revised (WAIS-R) Manual* (Wechsler, 1981) indicates that test-retest coefficients were determined by administering the WAIS-R twice, with 2 to 7 weeks intervening between testings, to 71 individuals in a 25-34 year age group, and to 48 individuals in a 45-54 year age group. At ages 25-34, test-retest means for the Verbal, Performance, and Full Scale IQs differed by about 3, 9, and 7 points, respectively. At ages 45-54, the differences between the IQ means for the three Scales were 3, 8, and 6 points, respectively. According to the WAIS-R manual, these gains revealed a practice effect when retesting an individual over a short time.

In their examination of the WAIS-R, Rapport, Brines, Axelrod, and Theisen (1997) examined the differential effects of practice as a function of Full Scale IQ over four administrations of the WAIS-R. Rapport et al. found that previous exposure to the WAIS-R dramatically altered performance. Of note was that there was a preponderance of improvement occurring in Performance versus Verbal IQ and the discrepancy between gain in Performance versus Verbal IQ was particularly great at the first retest. Rapport et al. indicated that the development of alternate forms for the Performance subtests would address practice effects associated with memory for specific items. However, such a solution would not address increased task familiarity for subtests such as Block Design. According to Rapport et al., the disproportionate gain in Performance IQ for the WAIS-R likely reflects a vulnerability of the Wechsler performance subtests associated with bonus points awarded for speed of solution. It was further suggested that future tests combine the development of a parallel form with restandardization of bonus points awarded for speed at retest.

The Wechsler Adult Intelligence Scale – Third Edition (WAIS-III; Wechsler, 1997) is a well-researched and widely used intelligence test for adults. The updated version of the *WAIS-III/WMS-III Technical Manual* (Wechsler, 2002) indicates that a number of studies have been conducted to provide evidence of the WAIS-III's reliability and validity as a comprehensive measure of adult intellectual functioning. Of note is that the WAIS-III included a new nonverbal subtest, Matrix Reasoning, which does not have time limits. Matrix Reasoning was added to the Performance scale and replaced Object Assembly, which relied heavily on bonus points for quick performance. Moreover, the *Technical Manual* indicates that the number of items with time-bonus points on the WAIS-III was decreased in the existing subtests. The aforementioned changes were made to decrease the reliance of the Performance scale on quick performance and subsequent bonus points.

In a study listed in the updated version of the *WAIS-III/WMS-III Technical Manual* (Wechsler, 2002), participants were tested twice, with a test-retest interval averaging 34.6 days. The data from the updated *Technical Manual* indicate that the mean retest scores are higher than the scores from the first testing. These differences, mainly due to practice effects, are about 2.0-3.2 points on the Verbal IQ (VIQ) score, 3.7-8.3 points on the Performance IQ (PIQ) score, and 3.2-5.7 points for the Full-Scale IQ (FSIQ) score. In the *Technical Manual*, it is noted that smaller gains in retest performance would be expected with test-retest intervals of relatively longer duration. In general, the older pooled age group (i.e., 55-89 years) showed smaller retest gains than did the 16-54 age group.

Groth-Marnat (2003) suggests that the reliabilities for the WAIS-III are generally quite high. However, Groth-Marnat indicates that, while test-retest reliabilities for the subtests of the WAIS-III indicate a high degree of temporal stability, there is still some degree of improvement on retesting because of practice effects. According to Groth-Marnat, the Full-Scale IQ on the WAIS-III was found to increase by 4.5 points, the Verbal IQ increased 2.4 points, and the Performance Scale increased by 6.5 points. Groth-Marnat notes that these increases are not only statistically significant, but may have clinical

significance when making inferences about the extent to which real improvement/deterioration has occurred for a particular examinee. According to Groth-Marnat, an examinee who has a Performance IQ increase of 6 points on retesting may not really be improving in his or her everyday functions, but is merely demonstrating practice effects. Kaufman and Lichtenberger (2002) suggest that a difference of 15 points (for ages 16 to 54) would be required to infer that there has been an actual improvement in abilities. Groth-Marnat notes that research with the WAIS-R indicates that these practice effects can occur up to nine months later even among head-injured patients.

Basso, Carona, Lowery, and Axelrod (2002) administered the WAIS-III to 51 participants at baseline and at an interval of either 3 or 6 months later. According to Basso et al., Full-Scale IQ (FSIQ), Verbal IQ (VIQ), Performance IQ (PIQ), Verbal Comprehension Index (VCI), Perceptual Organization Index (POI), and Processing Speed Index (PSI) scores improved significantly across time, whereas no significant change occurred on the Working Memory Index (WMI). Basso et al. add that, specifically, test scores increased approximately 3, 11, 6, 4, 8, and 7 points, respectively on the VIQ, PIQ, FSIQ, VCI, POI, and PSI for both groups and the degree of improvement was similar regardless of whether the inter-test interval was 3 or 6 months. Basso et al. conclude that these findings suggest that prior exposure to the WAIS-III yields considerable increases in test scores. As such, Basso et al. caution that users of the WAIS-III should interpret reevaluations across these intervals cautiously.

Less information on practice effects is available for the recently released Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV; Wechsler, 2008c) than for the WAIS-III. According to the *WAIS-IV Technical and Interpretive Manual* (Wechsler, 2008b), evidence of test-retest stability for subtest, process, and composite scores was obtained by administering the WAIS-IV twice, with test-retest intervals ranging from 8-82 days and a mean interval of 22 days. The test-retest reliability was estimated for four age bands (16-29, 30-54, 55-69, and 70-90) using Pearson's product-moment correlation. The average stability coefficients for all ages were calculated using Fisher's z transformation. The data suggest that the mean retest scores for all subtests are higher than the scores from the first testing. In general, test-retest gains are less pronounced for the Verbal Comprehension and Working Memory subsets than the Perceptual Reasoning and Processing Speed subtests. These results are generally consistent with those reported for the WAIS-III. As is true with the WAIS-III, test-retest reliabilities for the subtests and composite scores of the WAIS-IV show a high degree of temporal stability with some degree of improvement on retesting due to practice effects.

The WAIS-IV follows the tradition of the WAIS-III in its attempt to deemphasize time bonuses. For example, the *WAIS-IV Technical and Interpretive Manual* (Wechsler, 2008b) indicates that the number of Block Design items with time bonus points was reduced. Moreover, the use of time bonuses on Arithmetic was eliminated. Such modifications appear useful in reducing the tendency of performance items to be more susceptible than verbal items to practice effects.

According to the *WAIS-IV Administration and Scoring Manual* (Wechsler, 2008a), the shortest test-retest interval that will not result in significant practice effects on the WAIS-IV has not yet been determined. Of note is that the WAIS-IV contains three new subtests including Visual Puzzles, Figure Weights, and Cancellation. In the *Administration and Scoring Manual* it is noted that, if a retest of the WAIS-IV is necessary after a short time interval, supplemental subtests that were not used in the initial evaluation may be substituted for those administered in the initial evaluation. It is further indicated in the *Administration and Scoring Manual* that utilizing supplemental subtests is particularly important for subtests in the Perceptual Reasoning and Processing Speed scales of the WAIS-IV because they show the greatest practice effects after short time intervals. For the Perceptual Reasoning scale, at retest an examiner could substitute Figure Weights for Block Design and Picture Completion for Matrix Reasoning. Moreover, for the Processing Speed scale, Cancellation could be substituted for Coding or Symbol Search.

Practice Effects for the Stanford–Binet Intelligence Scales

The Stanford-Binet Intelligence Scales—Fifth Edition (SB5) is a widely used assessment tool for measuring intelligence (Roid, 2003). According to Roid, a key advantage of this intelligence test's most recent revision is that it includes improved low-end items for better measurement of young children or adults having mental retardation. Sbordone, Saul, and Purisch (2007) report that the range of the SB5 was expanded to allow the assessment of very low and very high levels of cognitive ability. Roid and Barram (2004) indicate that the practice effects on the SB5 were smaller than expected. For example, the nonverbal IQ of the SB5 showed shifts of only 2 to 5 points as compared to the 4 to 13 points on the Performance IQ of the Wechsler scales (i.e., the WAIS-III and WISC-III). Roid and Barram add that the lower shift, and thus practice effect, is even more notable given that the retest period for the SB5 was 5 to 8 days versus 23 to 35 days on average for the Wechsler scales.

Sbordone et al. (2007) agree with Roid and Barram (2004) that the practice effects for the SB5 are minimal. More specifically, Sbordone et al. suggest that the practice effects amounted to a Full-Scale IQ improvement of typically 2 to 4 points. Sbordone et al. emphasize that the SB5's minimal practice effects, minimal floor or ceiling effects, and excellent reliability make the test ideal for testing low functioning adults. Moreover, Roid and Barram indicate that the implications of the lower practice effect on the SB5 are that retesting can be done earlier on the SB5 than on other IQ batteries. More specifically, Roid and Barram report that retesting on the SB5 can occur in as little as 6 months rather than the typical one year delay.

Beta-III Practice Effects

The Beta III (Kellogg & Morton, 1999) is the revision of the Revised Beta Examination-Second Edition (Beta-II; Kellogg & Morton, 1978). The Beta-III is a group-administered, nonverbal test designed for use with individuals in the general population or with individuals who are non-English speakers, relatively illiterate, or who have language

difficulties. Since some individuals suspected of mental retardation struggle with language difficulties and/or literacy, it was deemed important to review a nonverbal intelligence test such as the Beta-III. Kellogg and Morton report that the reliability of scores from the Beta-III was assessed by the test-retest method. The sample consisted of 204 participants, tested between 2 and 12 weeks. The coefficients were displayed in three age-bands: 16-24, 25-54, and 55-89. The resulting coefficients were .87, .82, and .90, respectively, for the three age bands. When corrected for restriction of range, the coefficients were .91, .90, and .91, respectively. Kellogg and Morton note that the Beta-III IQ scores at the second testing are significantly higher than those at the first testing. According to Kellogg and Morton, for the overall sample, the discrepancy between the first and second testing is 7 points, which is mainly due to practice effect.

The sixteenth edition of the *Mental Measurements Yearbook* contains a review of the Beta-III by Soares. In the review, Soares suggests that it is possible that memory and practice effects might have influenced the Beta-III scores at the time of retest.

In summary, the available information indicates that practice effects are one of many factors that impact the stability of test scores. In general, the WAIS-R, WAIS-III, WAIS-IV, SB5, and Beta-III scores demonstrate adequate stability across time for all age groups. However, the available research indicates that most intelligence tests are influenced by practice effects. In the test-retest studies found in the test manuals for the WAIS-R, WAIS-III, WAIS-IV, SB5 and Beta-III, the mean retest scores of the subjects are higher than the scores from the time of the first testing. These differences can be attributed to practice effects. Somewhat surprisingly, such practice effects are less apparent on the SB5.

Larger Effects for Performance Items

Gains on retest are likely to be larger on performance items because examinees can develop problem-solving strategies that can be applied to the same or similar problems (Kaufman, 1994; Sattler, 2001). Sattler states that puzzles and block designs, for example, may be solved more easily on a repeated administration because the individual is familiar with the materials and can re-employ more efficiently problem-solving strategies that prove successful. Kaufman (1994) notes that, when considering the Wechsler subtests, Picture Arrangement and Object Assembly tend to produce large practice effects, and these tasks are consistently among the least reliable Wechsler subtests. Rapport et al. (1997) note that practice effects are particularly likely to occur on performance tests like Object Assembly, since such tests have an easily remembered single solution. Sattler indicates that another factor contributing to greater changes in performance scores is the importance of speed in determining an individual's score. Many tests are timed, and bonus points are awarded for correctly completing the items quickly.

The Importance of Timeframes in Examining Practice Effects

The test-retest studies on the reviewed intelligence tests involved relatively short timeframes. However, Basso et al. (2002) found that practice effects on the WAIS-III were still apparent at 3 and 6 months. Thus, practice effects for the reviewed tests are most evident in relatively short time frames. However, smaller gains in retest performance would be expected with test-retest intervals of relatively longer duration. As noted by Kaufman (1994), intervals that are relatively long, (e.g., over six months), permit the test taker to not remember most aspects of the test's content, which, in turn, reduces the magnitude of the practice effects. Given the available information, practice effects should be taken into account in the interpretation of intelligence test scores, especially when retesting occurs within six months to one year of the initial testing procedure. Practice effects could imply the presence of intellectual growth when no such growth has actually occurred.

The Impact of IQ on Practice Effects

The available research indicates that an individual's IQ at the time of initial testing has an impact on the extent to which practice effects will influence test scores at the time of retest. Shatz (1981), for example, suggested that individuals with cerebral dysfunction are not expected to show practice effects with a single retesting. According to Mitrushina and Satz (1991), individuals with dementia can be expected to show further deterioration rather than improvement on follow-up evaluations of cognitive functioning. However, Mitrushina and Satz note that this assumption has some practical limitations stemming from repeated use of the same cognitive measures on initial and follow-up evaluations. According to Mitrushina and Satz, due to the influence of practice effects, an increase at retest does not necessarily reflect improvement in cognitive functioning, and retaining the same or similar scores at retest would not necessarily indicate a lack of decline because practice effects may counterbalance cognitive decline.

Rapport et al. (1997) found that practice effects appear to have less of an impact on individuals with lower IQs than for individuals with higher IQs. More specifically, differential effects of practice over four administrations of the WAIS-R were examined by Rapport et al. as a function of Full-Scale IQ at initial testing. In Rapport et al.'s study, participants were tested at 2-week intervals and a repeated measures analysis of variance indicated that Average and High-Average groups made greater gains across retest intervals than did the Low-Average group.

A prosecutor might argue that courts should not be concerned with practice effects among individuals who are thought to be mentally retarded since being administered an intelligence test on multiple occasions has less of an impact on their scores than for individuals with higher IQs. Although individuals with low IQs might not demonstrate as much benefit from practice effects as individuals with higher IQs, this does not change the fact that individuals with low IQs might show enough of an increase at retest to place them above the 70 IQ threshold some states have established as representing mental retardation.

Frequency of Re-evaluation

In the aforementioned research by Rapport et al. (1997), across groups, gains on the WAIS-R were greater at the first retest than at the second or third retest. This finding flies in the face of the logical assumption that performance should improve over time as the areas being assessed become increasingly familiar. Theisen, Rapport, Axelrod, and Brines (1998) examined practice effects over four administrations of the immediate (I) and delayed (II) portions of three subtests of the Wechsler Memory Scales-Revised (WMS-R). The greatest increase in scores occurred at the first retest session, whereas increases of smaller magnitudes occurred at Sessions 3 and 4. Similarly, Ivnik et al. (1999) administered the WAIS-R as a part of a larger test battery yielding five cognitive factors. Ivnik et al. examined cognitively normal older persons (over age 54) at test intervals of 1 and 2 years. For persons tested at 1- to 2-year intervals, practice effects were demonstrable only between the first and second assessments. Thus, on a variety of cognitive measures including the Wechsler scales, after the second assessment, practice effects tend to level off.

The Age Effect

Research has demonstrated that practice effects on measures of general intelligence change as people grow older. For example, as noted in the updated *WAIS-III/WMS-III Technical Manual* (Wechsler, 2002), the 55-89 year-old age group showed smaller retest gains on the WAIS-III than did the 16-54 age group. Moreover, in his examination of different studies that assessed normal elderly adults, Shatz (1981) noted that gain on retest did not exceed 2 IQ points, which is lower than the gain of 5 Full-Scale IQ points at retest for adults noted by Matarazzo et al. (1980). Dikmen, Heaton, Grant, and Temkin's (1999) research also shows that, on tests of cognitive ability, younger participants tend to benefit more from practice.

The age span reviewed by Shatz (1981) was limited to ages 19 through 70. Mitrushina and Satz (1991) examined the magnitude of practice effects in repeated administration of neuropsychiatric measures, including the WAIS-R Performance Intelligence Quotient (PIQ), which tap different cognitive domains in 122 normal elderly subjects between the ages of 57 and 85. The subjects were evaluated over three annual testing probes. Mitrushina and Satz found that individuals age 57 to 65 demonstrate remarkable improvement on the retest with WAIS-R Performance subtests that can be attributed to practice effects while 66 to 75-year-old people are less likely to show a practice effect. Mitrushina and Satz also found that people over the age of 75 do not benefit from previous exposure to the tests and demonstrate decline at retest. Mitrushina and Satz further suggest that a practice effect is not evident in Verbal subsets for any elderly age group. Mirushina and Satz's study indicates that practice effects should not be considered independent of the population being examined.

Ronnlund and Nilsson (2006) examined aging patterns in the WAIS-R Block Design Test (BDT) cross-sectionally and longitudinally. The cross-sectional analyses indicated

a gradual age-related deterioration from 35 to 85 years. The longitudinal data, on the other hand, showed stable performance from age 35 to 55, even when minor practice effects were adjusted for, and decline past age 55. Of note was that the longitudinal data were suggestive of little, if any, decrement in mean-level performance before age 60.

Implications for Practice

The aforementioned review suggests that courts ordering follow-up cognitive capacity evaluations of individuals on death row should not receive a follow-up examination with the same intelligence test within a relatively short time frame. This review further indicates that, even for longer time frames such as 6 months to one year, practice effects can still have an impact on the defendant's IQ scores. A general rule of thumb is that most intelligence tests, with the possible exception of the SB5, should not be readministered within one year of the most recent test administration. As noted, the available research suggests that, even though the SB5 can be readministered earlier than other measures of intelligence, clinicians should still wait at least 6 months. Clinicians who are performing follow-up examinations within 6 months of the initial test administration should strongly consider the use of a different IQ test than was used at the time of the initial assessment of the defendant.

It was previously indicated that in some states specific cutoff scores are utilized to determine if a person should be considered mentally retarded. It is possible that knowledgeable prosecutors in such states might capitalize on information concerning practice effects by having defendants evaluated multiple times within relatively short time-frames. Clinicians considering performing IQ testing in capital cases in states with strict IQ cutoffs need to carefully consider the ethical implications of administering the same IQ test within a short time-frame. Duvall and Morris (2006) suggest that psychologists refuse appointments to provide assessments in those states whose statutes clearly violate sound psychometric practices. Although refusing an appointment is certainly an option worth considering, another possible solution would be to utilize a different intelligence test in circumstances when there is a requirement to administer IQ testing multiple times within a short period of time.

As noted, the *WAIS-IV Administration and Scoring Manual* (Wechsler, 2008a) suggests that clinicians can use the supplemental subtests of the WAIS-IV at the time of retest to substitute for certain subtests used in the initial evaluation. The *Administration and Scoring Manual* indicates that such a substitution method is particularly important for subtests in the Perceptual Reasoning and Processing Speed scales because they show the greatest practice effects after short time intervals. One problem with the substitution approach, however, is to what degree substituting alternate subtests will affect the Full-Scale IQ score. The *Administration and Scoring Manual* indicates that no more than two substitutions are allowed when deriving the Full-Scale IQ score. Moreover, each index score may only include one substitution. The *Administration and Scoring Manual* acknowledges that substitution introduces the risk of increased measurement error. Based on the limited number of allowed substitutions and the possibility of

measurement error, utilizing an entirely different IQ test at retest appears to be a more optimal solution to reduce the influence of practice effects.

Clinicians should communicate IQ test results in terms of confidence intervals. Researchers such as Matarazzo and Prifitera (1989) note that the magnitudes of the standard errors of measurement associated with intelligence tests such as the Wechsler scales highlight the risk of using a scaled score from one administration of an intelligence test. Matarazzo and Prifitera advocate that clinicians consider a band of scores extending two standard errors above and below the obtained score as a practical method of taking into account the standard error of measurement. Communicating IQ scores in terms of confidence intervals demonstrates to the court that an individual's IQ score varies and cannot accurately be characterized by a single number.

Iverson (2001) and Chelune (as cited in Dawes & Senior, 2001) have devised methods for evaluating meaningful change in test scores with regard to the WAIS-III. According to Dawes and Senior, the approaches described by Chelune and Iverson compute critical values above which systematic change, not attributable to practice, is detected. Dawes and Senior proposed a method that uses the standard error of prediction to estimate a 90% confidence interval band around the individual's predicted true score. According to Dawes and Senior, this method makes no assumptions regarding the magnitude of systematic causes of change such as practice and the confidence band computed is based solely upon the psychometric characteristics of the test. Dawes and Senior note that the confidence band indicates the range in which test scores are expected to vary as a consequence of their reliability and variability. Scores that fall outside of the band are interpreted as resulting from systematic change.

The aforementioned approaches by Iverson (2001), Chelune (as cited in Dawes & Senior, 2001) and Dawes and Senior generate confidence bands that can be useful in detecting systematic changes in test scores over time. Dawes and Senior point out that, unlike the approaches by Iverson and Chelune, which explicitly assume that practice effects must have influenced test scores in each and every application, the standard error of prediction approach has the advantage of utilizing psychometric characteristics that assume no systematic influences resulting in a measure that is much more sensitive to meaningful change. The confidence bands proposed in all three of these methods provide a means by which clinicians can limit the extent to which practice effects or other systematic changes affect the measurement of IQ scores over time. Such confidence bands would appear to be particularly useful in capital cases given the high stakes involved in such evaluations. Future research is needed to determine the applicability of the aforementioned confidence band approaches to the WAIS-IV and other commonly administered intelligence tests.

It is important for clinicians examining a defendant's intellectual functioning at retest to carefully consider external variables such as intelligence and age. As noted, the available research suggests that differential amounts of practice effects should be expected as a function of the defendant's age (Dikmen et al., 1999; Mitrushina & Satz, 1991; Ronnlund and Nilsson, 2006; Shatz, 1981) and competency level (Ivnik et al.,

1999; Mitrushina and Satz, 1991; Rapport et al., 1997; Shatz, 1981). Moreover, the research shows that practice effects appear to level off after the first retest (Ivnik et al., 1999; Rapport et al., 1997; Theisen et al., 1998). Based on this research, examiners would expect younger, more competent individuals to perform relatively better at the time of the first retest. Although the impact of these variables is difficult to measure with any degree of precision, at minimum clinicians need to communicate to the court in their evaluation reports the potential influence of age, competency, and the number of times the defendant has been retested on the defendant's test scores.

Before making conclusions concerning whether or not a particular defendant has mental retardation, careful record reviews are important. Obtaining appropriate collateral information is critical in establishing whether or not a person is mentally retarded because most definitions of mental retardation (e.g., the DSM-IV) suggest that the deficits need to be apparent at a relatively young age, typically prior to age 18. As such, clinicians should compare the current obtained test results to the available prior records and be sure to explain any score differences. Based on the influence of practice effects, it is possible for a defendant who has historically scored below 70 on prior examinations of intelligence to score above 70 if he or she is retested on the same IQ measure within a short time frame. Examining prior IQ results, then, can provide clinicians with a clearer picture of the accuracy of a defendant's IQ score at retest.

Another critical aspect of determining a person's IQ is through an examination of adaptive behavioral functioning. Again, most definitions of mental retardation, such as the definition in the DSM-IV, indicate that, to be considered mentally retarded, a person must have concomitant deficits in adaptive behavior. As such, it is important for clinicians to perform a thorough examination of the defendant's adaptive behavioral functioning using professionally valid and appropriate instruments. Many adaptive behavioral instruments rely on third-party ratings (e.g., Frumkin, 2006; Yalon-Chamovitz & Greenspan, 2005), which can introduce a potential for bias and inaccuracy. Another problem with the available adaptive-behavioral instruments cited by Frumkin is that they ask about activities that most defendants do not typically engage in. Moreover, Frumkin indicates that most of the available adaptive-behavioral instruments depend on ratings by informants who are not aware of the defendant's current functioning and none include norms developed to compare a defendant's scores with a correctional population. A full discussion of the difficulties associated with measuring adaptive-behavioral functioning is beyond the scope of this article. What is clear, however, is that, despite problems inherent in accurately measuring it, adaptive-behavioral functioning is important because it provides a more holistic examination of the defendant and minimizes sole reliance on IQ test scores in determining if a person meets criteria to be considered mentally retarded.

Based on the reviewed information, when clinicians assess the intelligence of defendants involved in capital cases, the following guidelines are recommended:

1. Determine what intelligence test was used in the initial assessment of intelligence.

2. If possible, use a different intelligence test at the time of retesting, particularly if the time frame is less than one year.

3. If the same intelligence test is used at retest, consider using supplemental subtests to substitute for subtests used in the initial evaluation.

4. Utilize adaptive behavioral testing and a careful review of the available collateral information in making conclusions about mental retardation.

5. Communicate IQ results in terms of confidence intervals and clarify possible practice effects in the report to the court.

6. Consider to what extent the examinee's initial IQ, age at the time of initial testing, and the number of times the examinee has been tested impact his or her current test results and communicate the influence of these variables to the court.

7. Consider refusing appointments to provide assessments of intelligence in states with statutes that clearly violate sound psychometric practices.

In summary, this review demonstrates that practice effects are often associated with artificial increases in intellectual test scores, particularly over relatively short time frames. Moreover, these increases can have a significant impact on whether or not a person is considered mentally retarded. Because mental retardation directly impacts the court's determination of whether or not the defendant is eligible to be put to death, clinicians assessing mental retardation in capital cases are ethically obligated to consider practice effects. Adhering to the aforementioned guidelines provides clinicians with practical methods to eliminate or minimize the extent to which practice effects impact whether or not a defendant is considered mentally retarded.

Received on August 14, 2009; accepted on September 29, 2009; final version received on October 9, 2009.

References

American Psychological Association. (2002). American Psychological Association ethical principles of psychologists and code of conduct. Retrieved from http://www.apa.org/ethics/code2002.html

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

Basso, M. R., Carona, F. D., Lowery, N., & Axelrod, B. N. (2002). Practice effects on the WAIS-III across 3 and 6-month intervals. *The Clinical Neuropsychologist, 16(1),* 57-63.

Bobby v. Bies, 129 S. Ct. 2145 (2009).

Cunningham, M. D., & Goldstein, A. M. (2003). Sentencing determinations in death penalty cases. In A. Goldstein (Ed.), *Forensic psychology* (Vol. 11, pp. 407-436). New York: John Wiley.

Dawes, S., & Senior, G. (2001, October). *Lessons from the WAIS-III: 3 issues when retesting with the WAIS-III.* Poster session presented at the 7th annual conference of the APS College of Clinical Neuropsychologists. Melbourne, Australia. Retrieved from http://www.usq.edu.au/users/senior/Posters/WAIS-III%20Retest.htm.

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. T. (1999). Test–retest reliability and practice effects of Expanded Halstead–Reitan. *Journal of the International Neuropsychological Society*, *5*, 346–356.

Duvall, J. C., & Morris, R. J. (2006). Assessing mental retardation in death penalty cases: Critical issues for psychology and psychological practice*. Professional Psychology: Research and Practice, 37(6),* 658–665. doi: 10.1037/0735-7028.37.6.658

Frumkin, I. B. (2006). Challenging expert testimony on intelligence and mental retardation. *The Journal of Psychiatry & Law, 34*, 51-71.

Groth-Marnat, G. (2003). *Handbook of psychological assessment (4th ed.)*. Hoboken, New Jersey: Wiley & Sons.

Ivnik, R. J., Smith, G. E., Lucas, J. A., Petersen, R. C., Boeve, B. F., Kokmen, E., & Tangalos, E. G. (1999). Testing normal older people three or four times at 1 to 2 year intervals: Defining normal variance. *Neuropsychology*, *13*, 121-127.

Kaufman, A. S. (1994). Practice effects. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence: Vol. 2*. (pp. 828-833). New York: Macmillan.

Kaufman, A. S., & Lichtenberger, E. O. (2002). *Assessing adolescent and adult intelligence*. Boston: Allyn & Bacon.

Kellogg, C. E., & Morton, N. W. (1978). *Revised Beta Examination—Second Edition*. San Antonio, TX: The Psychological Corporation.

Kellogg, C. E., & Morton, N. W. (1999). *Beta-III manual*. San Antonio, TX: The Psychological Corporation.

Matarazzo, J. D., Carmody, T. P., & Jacobs, L. D. (1980). Test-retest reliability and stability of the WAIS: A literature review with implications for clinical practice. *Journal of Clinical Neuropsychology*, *2*, 89–105.

Matarazzo, J. D., & Prifitera, A. (1989). Subtest scatter and premorbid intelligence: Lessons from the WAIS-R standardization sample. *Psychological Assessment, 1,* 186-191.

Mitrushina, M., & Satz, P. (1991). Effect of repeated administration of a neuropsychological battery in the elderly. *Journal of Clinical Psychology 47(6)*, 790-801.

Murphy v. Oklahoma, Okla. Crim. App. LEXIS 37 (Sept. 4, 2002).

Rapport, L. J., Brines, D. B., Axelrod, B. N., & Theisen, M. E. (1997). Full scale IQ as mediator of practice effects: The rich get richer. *The Clinical Neuropsychologist*, *11(4),* 375-380.

Roid, G. H. (2003). *Stanford–Binet intelligence scales, fifth edition, technical manual.* Itasca, IL: Riverside.

Roid, G. H., & Barram, R. A. (2004). *Essentials of Stanford-Binet intelligence scales (SB5) assessment*. John Wiley and Sons.

Ronnlund, M., & Nillson, L.G. (2006). Adult life span patterns in WAIS-R Block Design performance: Cross sectional versus longitudinal age gradients and relations to demographic factors. *Intelligence, 34*, 63-78.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications (4th ed.)*. San Diego: Jerome M. Sattler Publisher, Inc.

Savage, D. G. (2007, June 11). IQ debate unsettled in death penalty cases. *Los Angeles Times*. Retrieved from http://www.probono.net/deathpenalty/news/article.150133-IQ_Debate_Unsettled_in_Death_Penalty_Cases.

Sbordone, R. J., Saul, R. E., & Purisch, A. D. (2007). *Neuropsychology for psychologists, health care professionals, and attorneys (3rd ed.)*. CRC Press.

Shatz, M. W. (1981). WAIS practice effects in clinical neuropsychology. *Journal of Clinical Neuropsychology, 3*, 171-179.

Soares, L. M. (n.d.). [Review of the test Beta-III]. In *the sixteenth mental measurements yearbook*. Retrieved from EBSCOHost Mental Measurements Yearbook database.

Theisen, M. E., Rapport, L. J., Axelrod, B. N., & Brines, D. B. (1998). Effects of practice in repeated administrations of the Wechsler Memory Scale—Revised in normal adults. *Assessment*, *5,* 85-92.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*. New Jersey: Pearson.

Wechsler, D. (1955). *Wechsler Adult Intelligence Scale*. New York: Psychological Corporation.

Wechsler, D. (1981). *WAIS-R manual: Wechsler adult intelligence scale-revised*. New York: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (2002). *WAIS-III/WMS-III technical manual, updated*. San Antonio, TX: Harcourt Brace.

Wechsler, D. (2008a). *WAIS-IV administration and scoring manual*. San Antonio, TX: Pearson.

Wechsler, D. (2008b). *WAIS-IV technical and interpretive manual*. San Antonio, TX: Pearson.

Wechsler, D. (2008c). *Wechsler Adult Intelligence Scale* (4th ed.). San Antonio, TX: Pearson.

Yalon-Chamovitz, S., & Greenspan, S. (2005). Ability to identify, explain and solve problems of everyday tasks: Preliminary validation of a direct video measure of practical intelligence. *Research in Developmental Disabilities, 26,* 219–230. doi:10.1016/j.ridd.2004.08.002