APPLIED NEUROPSYCHOLOGY, *16*: 214–222, 2009

ISSN: 0908-4282 print/1532-4826 online
DOI: 10.1080/09084280903098786

Psychology Press
Taylor & Francis Group

Serious Problems with the Mexican Norms for the WAIS-III when Assessing Mental Retardation in Capital Cases

Hoi K. Suen
Educational Psychology Program, Pennsylvania State University, University Park, Pennsylvania

Stephen Greenspan
Department of Psychiatry, University of Colorado, Denver, Colorado

A Spanish-language translation of the Wechsler Adult Intelligence Scale-III (WAIS-III), normed in Mexico, is sometimes used when evaluating Spanish-speaking defendants in capital cases in order to diagnose possible mental retardation (MR). Although the manual for the Mexican test suggests use of the U.S. norms when diagnosing MR, the Mexican norms—which produce full-scale scores on average 12 points higher—are sometimes used for reasons that are similar to those used by proponents for "race-norming" in special education. Such an argument assumes, however, that the Mexican WAIS-III norms are valid. In this paper, we examined the validity of the Mexican WAIS-III norms and found six very serious problems with those norms: (1) extremely poor reliability, (2) lack of a meaningful reference population, (3) lack of score normalization, (4) exclusion of certain groups from the standardization sample, (5) use of incorrect statistics and calculations, and (6) incorrect application of the true score confidence interval method. An additional problem is the apparent absence of any social policy consensus within Mexico as to the definition and boundary parameters of MR. Taken together, these concerns lead one to the inescapable conclusion that the Mexican WAIS-III norms are not interpretable and should not be used for any high-stakes purpose, especially one as serious as whether a defendant should qualify for exemption against imposition of the death penalty.

Key words: *Atkins hearing, death penalty, mental retardation, WAIS-III*

INTRODUCTION

This paper grew out of a pre-trial "Atkins" (death penalty exemption due to possible mental retardation) hearing in which the two of us were defense experts. The case involved a 37-year-old man who immigrated recently to the United States illegally from Mexico and was indicted for murdering a woman with whom he was having an affair. Because the defendant had very limited English skills, a Spanish-speaking neuropsychologist hired by the defense administered a Mexican-Spanish version of the Wechsler Adult Intelligence Scale-III (WAIS-III), the Escala *Wechsler de Inteligencia Para Adultos-III*, published in Mexico (Tulsky & Zhu, 2003). The Mexican test uses all of the items from the original WAIS-III but made minor changes to some of the instructions and item hierarchy in order to be more sensitive to cultural factors. The Mexican WAIS-III provides two sets of norms—the original U.S. WAIS-III norms and newly developed Mexican norms—and users are given the option of using one or the other. The defense psychologist chose to use the U.S. norms and

Address correspondence to Hoi K. Suen, Pennsylvania State University, 103 CEDAR Building, University Park, PA 16802. E-mail: HoiSuen@psu.edu

obtained a full-scale IQ score of 66, which is under the 70–75 ceiling for diagnosing mild mental retardation (MR), even without doing a "Flynn" adjustment for norm-obsolescence. Another Spanish-speaking neuropsychologist was hired by the court, and he did two things: He re-scored the defense psychologist's test results, using the Mexican norms, and he re-administered the Mexican WAIS-III and scored it both ways. The obtained scores and confidence intervals are depicted in Table 1.

The two psychologists obtained full-scale IQ scores that were quite close, with the defense psychologist's prior scores (perhaps because of the Practice Effect) being four points lower using the U.S. norms (66 and 70, respectively) and two points lower using the Mexican norms (79 and 81, respectively). As can be seen, the Mexican norms provided results that were 13 and 11 points higher, placing the defendant well above the 70–75 ceiling. The court-appointed psychologist, who ended up testifying for the prosecution, argued along with the prosecutor that with a recently arrived Mexican national, even one residing (and possibly committing a crime) in the United States, it is a no-brainer that one should use the Mexican norms. They also used a kind of race-norming argument, similar to the ones used by Mercer (1988) to reduce over-assignment of poor minority children to special education, and by Heaton, Taylor, and Manley (2003) to reduce over-diagnosis of dementia in African Americans. In this view, the Mexican norms, which ruled out MR for this defendant, were more appropriate because he was more similar culturally and educationally to the Mexican population mean and thus looked less deficient when compared to that population.

The defense attorneys' position on these issues was that sufficient cultural sensitivity was shown by administering a Spanish-language version of the test. In their view, the appropriate standard to use when deciding an Atkins claim for a crime committed in the United States is to consider how the defendant functions relative to the U.S. population rather than to the population of Mexico, a country which has neither the death penalty nor for that matter any apparent official definition of MR. With respect to the race-norming argument, the defense position was that this is a controversial, and possibly illegal, practice. Again, their view was that a diagnosis of MR reflects very low functioning relative to the general population and that by correcting scores upwards to re-norm them within a particular subgroup, one would be cutting off many (perhaps most) deserving individuals from the supports and protections to which they are legally entitled. An irony about race-norming in death penalty cases is that a score adaptation mechanism intended for the educational and human services benefit of poor minority individuals would now have the effect, if adopted, of increasing their already too-high representation on death rows.

TABLE 1
WAIS-III Scores and Confidence Intervals Using U.S. and Mexican Norms

IQ scores	*Using U.S. norms (95% CI)*	*Using Mexican norms (95% CI)*
Defense psychologist	66 (63–71)	79 (65–103)
Court-appointed psychologist	70 (67–75)	81 (67–104)

The second author (SG) was hired by the defense to be a teaching expert on MR, with particular emphasis on adaptive behavior. Although not an expert on psychometrics, he was struck by the huge confidence intervals (38 and 37 at the 95% level) for the two scores obtained using the Mexican norms and by the highly asymmetrical nature of those confidence intervals. The apparent poor reliability of the Mexican scores was attributed by the Mexican norms developers (in a translation obtained for the Mexican technical manual) to a smaller norming sample (970 as opposed to 2,450 for the U.S. norms), an explanation which made no sense. He was also struck by a statement in the manual (Tulsky & Zhu, 2003) to the effect that because of limited resources, the publisher was unable to construct the norming sample in a manner that adequately represented the Mexican population. For example, there were no subjects with MR in the standardization sample, which was a reason given in the manual for the recommendation that the U.S. norms should be used for the purpose of diagnosing MR.

The earlier-stated prosecution's legal and philosophical arguments in favor of using the Mexican norms may be tenable—the potential problems of race-norming and the lack of Mexican MR standard notwithstanding—only if the Mexican norms turn out to be generally adequate. If they turn out to be grossly inadequate, then the IQ scores produced by the Mexican norms cannot be relied upon for any high-stakes decision, let alone one as serious as a request for death penalty relief. Because of this concern, the second author suggested that the defense engage the services of a psychometrics expert, which turned out to be the first author (HS). After carefully studying the material in the Mexican technical manual (Tulsky & Zhu, 2003), discussing various issues with the second author and other experts, and consulting a number of references, the first author produced a report which argued strongly that the Mexican WAIS-III norms are invalid and are not interpretable. The reasons spelled out in the report are discussed in the following sections.

PROBLEM ONE WITH THE MEXICAN NORMS: MUCH TOO LOW RELIABILITY

The technical manual for the Mexican WAIS-III test reports a Cronbach Alpha reliability coefficient value

of .8613 for the overall IQ scores based on the Mexican norms. (This was reported as .8677 at a different point in the manual, perhaps referring to a different sample.) The reported reliability coefficient value of either .8613 or .8677 would indicate that the scores are quite reliable, with relatively little measurement error. Yet, this relatively high reported reliability coefficient value is inconsistent with the reliability value otherwise implicit in the confidence interval information provided for the scores reported for the defendant.

To explore this inconsistency, we need to first review the relationship between reliability and confidence interval. The range of values for a confidence interval is related to the value of the corresponding reliability coefficient via the standard error of measurement (SEM) as follows:

$$95\%\,\text{confidence interval} = \hat{\tau} \pm 1.96\sigma_e \qquad (1)$$

where $\hat{\tau}$ is an estimated true score, and σ_e is the SEM. There are a number of methods available to estimate SEM values. For WAIS-III scores, a ''true score confidence interval'' approach proposed by Glutting, McDermott, and Stanley (1987) is used. Under this approach, there are three alternative procedures for the estimation of SEM. These are the Stanley method, the Lord and Novick method, and the Nunnally method. These three methods are shown below:

$$\text{Stanley Method}\ \hat{\sigma}_e = \sigma_x\sqrt{1-\hat{\rho}_{xx}}\cdot(\hat{\rho}_{xx}) \qquad (2)$$

where σ_x is the standard deviation of the observed scores and $\hat{\rho}_{xx}$ is the estimated reliability coefficient.

$$\text{Lord and Novick}\ \hat{\sigma}_e = \sigma_x\sqrt{(1-\hat{\rho}_{xx})\hat{\rho}_{xx}} \qquad (3)$$

and

$$\text{Nunnally Method}\ \hat{\sigma}_e = \sigma_x\sqrt{1-\hat{\rho}_{xx}} \qquad (4)$$

The Stanley method is the specific method used for the original U.S. WAIS-III scores. The method used for the translated Mexican WAIS-III scores is not reported.

Given the confidence interval values reported by the two psychologists shown in Table 1, it can be deduced based on Equation 1 that the SEM value for the Mexican scores is about 9.95. If the Cronbach alpha reliability coefficient value of .8613 (or .8677) reported in the Mexican WAIS-III technical manual is indeed correct, we ought to be able to arrive at this exact same SEM value of 9.95 based on the Cronbach alpha value using the mathematical relations described in either Equation 2, 3, or 4, depending on which exact method was used by the Mexican test developer. Yet, when the Cronbach alpha value of .8613 was applied, the value of SEM obtained is either 4.83 (Stanley method), 5.20 (Lord and Novick method), or 5.61 (Nunnally method)—in all cases only about half the magnitude of the 9.95 obtained from the confidence interval side of the equation. The differences are even slightly larger when the Cronbach alpha value of .8677 is applied.

In fact, given that IQ scores have predetermined standard deviation values of 15, under no circumstance can an SEM value of 9.95 and a Cronbach alpha value of .8613 or .8677 coexist, even taking into account minor calculation or rounding errors. Therefore, there is contradictory information regarding the reliability and confidence interval of the Mexican scores.

The only situation under which the SEM value can be 9.95 is when the Nunnally method is used and the actual reliability coefficient value is .56 instead of the reported Cronbach alpha values of .8613 and .8677. For test scores that are used in a manner that can lead to serious consequences, we generally demand very high reliability coefficient values that are close to 1.0. A reliability coefficient value of .56 would be generally considered unacceptably low. If the correct reliability coefficient of the Mexican scores is indeed .56, we can consider the scores unreliable.

Given the conflicting information, we do not know which is the real situation with the Mexican scores. However, given that the decision regarding the MR status of the defendant was partly based on the reported IQ scores and its 95% confidence intervals, the corresponding reliability coefficient of .56 or less is effectively the operating reliability for the purpose of the hearing.

PROBLEM TWO WITH THE MEXICAN NORMS: THE LACK OF A MEANINGFUL REFERENCE POPULATION

IQ scores are norm-referenced scores. A score is interpreted as an indicator of how well a person performs on the test relative to the performance of other individuals in a particular reference population. The scoring metric used has specific meaning in relation to this particular reference population. For example, an IQ score of 100 means the person scores higher than 50% of the reference population, and an IQ score of 115 means the person scores higher than about 84% of the reference population. Similarly, an IQ score of 70 means the person scores higher than about 2.3% of the reference population. Therefore, IQ scores need to have a clearly identified and well-represented reference population to be interpretable. Norms need to be established based on data from a clearly representative sample of that reference population. If either the reference population is not well defined or the standardization sample is not representative of this reference population, we would

have no basis to interpret the scores, and the scores would essentially not be interpretable.

According to the technical manual for the U.S. norms (The Psychological Corporation, 1997), the reference population for the U.S. norms is clearly defined as the adult population, ages 16 to 89, of the United States. The standardization or norming sample was carefully drawn using a scientifically sound proportional stratified sampling process. This process was designed to ensure that the standardization sample used was representative of the U.S. population in terms of sex, race/ethnicity, educational level, and geographic region, based on the distribution of these characteristics in the United States as was reported by the U.S. Census Bureau in 1995. Note that the 1995 Census data were the most up-to-date data available at the time of test development for the U.S. norms. The detailed distribution of sample subjects on each of these characteristics by age group was compared against the detailed distribution of U.S. adults on the same characteristic in the Census report to ensure precise representation (see Figures 2.2 and 2.3 and Tables 2.2 through 2.12 in the U.S. Technical Manual by the Psychological Corporation, 1997). The result is that the standardization sample is a representative microcosm of the U.S. population with respect to these characteristics, and the reference population for score interpretations is thus known and clearly delineated.

In contrast, the Mexican reference population is ill defined and the standardization sample is not representative. The following are excerpts from the court translation of the ''Methodology'' section of the Mexican WAIS-III technical manual:

> Regarding the characteristics of the sample for the standardization on a national level for Mexico, this was done taking into account both that it is technically impossible to get a sample by chance, based on social status of the Mexican population in general, as well as the need to follow and meet the resources and limitations of this research. Therefore, we did a sample based on conventional criteria according to the measurement theory of intelligence. It was decided originally of 1,000 Mexicans, male and female, with a defined level of education based on a high and low education criteria, and coming from four geographical regions of the country—the center, north, west, and southeast. To avoid differences it was decided to do a manipulation a priori of the variables that could prejudge the results, by instances, we tried to find the balance regarding gender, region of the country, age group, as well as academic education level of all the participants.
>
> ...This way, the definitive sample was formed by Mexicans from urban zones belonging to diverse social statuses. And each own, a relative balance based on age group (around 10% by social status) and by region (closer to 25% for each one of them). However, it is important to clarify that we detected variation corresponding to the over-representation of Mexicans of high level of education who finished high school or higher, (57/43%, $\chi^2 = 4.86$; $p \geq .001$).

Several problems become apparent in these paragraphs. First, the passage appears to indicate that it is not possible to collect a probability sample due to a limitation of resources. A probability sample (translated as ''sample by chance'' above) is the hallmark of scientific sampling, which enables the researcher to remove systematic bias from the sample. Without such a sampling process, the sample is likely to be biased. Second, the test developer decided to use the same set of characteristics as those used in the U.S. standardization sample to draw the Mexican sample. Yet no information on the distribution of these characteristics in Mexico was presented; no Mexican census or any population count was referenced, nor is there evidence that the actual Mexican population distribution on these characteristics was ever considered when drawing the standardization sample. Without knowledge of the distribution of the Mexican population on these characteristics, there is no way to ensure that the sample is representative of the Mexican adult population as there is no basis to determine proportions in the stratified sample.

The Mexican test developer reported that, ''we tried to find the balance regarding gender, region of the country, age group, as well as academic education level of all the participants'' and ''closer to 25% for each of [the four regions].'' It appears that the developer's goal was to draw samples so that there were about equal numbers of males and females, equal numbers of subjects from each of the four geographic regions of Mexico, and so on, and the developer described this as ''to find the balance.'' It is almost certain that this ''equal proportion'' method has produced a disproportional stratified sample that is not representative of the Mexican adult population, because it is highly improbable that there are exactly equal numbers of adults living in the four geographic regions of Mexico and exact numbers of adults with high and low levels of education.

In fact, the Mexican test developer found that 57% of the subjects in the standardization sample had a high level of education, and only 43% had a low level of education. The author described this as an ''over-representation of high level of education'' and provided the results of a chi-square test as proof of this over-representation. Although not stated, the most probable chi-square test used was a ''goodness-of-fit test.'' For this test, the author would need to specify the expected population distribution between high and low levels of education or the default would be 50/50. Since no expected population distribution is described by the

author, it would mean that a default of a 50/50 division was used. But such a division is meaningful only if there is indeed a 50/50 division of individuals with high and low educational levels among Mexican adults. What constitutes high and low educational level for the standardization sample is not clearly stated. If it was consistent with a definition of ''scholarity'' used for the sample in an earlier pilot study (described/translated as ''preliminary validation'' in the translation), a high level is someone with a high school diploma or above, and a low level is someone with three years of junior high school education or less. There is no evidence to either indicate or support the notion that the proportional division of adults with high school diplomas and above and those with three years of junior high school education and less in Mexico is exactly 50/50.

Since the developer provided no data about the adult population of Mexico, we are unable to determine whether the sample used was representative of this population at all. It is reasonable to deduce that it is extremely unlikely that there just happens to be exactly the same proportion of adults residing in each of the four regions used in the sample and exactly the same numbers of adults with high and low educational levels in the Mexican population.

If the developer had indeed consulted some unreported information about the distribution of the Mexican adult population and had determined based on that information that individuals with a high level of education are over-represented in relation to that distribution, it would still be possible to correct for that over-representation using statistical weighting techniques. There is no evidence that weighting methods were used at all. Instead, the developer opted to caution the user on the lack of representation of the sample. Again, according to the translated manual, the authors of the Mexican manual stated in the ''Methodology'' section: ''This clarification of the sample characteristics allows us to establish the limitations of the same and to determine the type of person that is susceptible of being evaluated fairly with the Mexican version.'' In spite of this recognition of the limitations of the sample, the authors provided no indication as to what type of person can be evaluated fairly with the Mexican version.

In summary, the distribution of characteristics of the Mexican adult population is unknown and unreported. The standardization sample is unlikely to be representative of this unknown Mexican population. Therefore, there is no clear reference population for the interpretation of the Mexican scores. For example, a Mexican IQ score of 115 is supposed to be indicative of a person who outperforms 84% of *some* unknown Mexican Spanish-speaking population or sub-population; but we have no means to know exactly *what* Mexican Spanish-speaking population or subpopulation. The only thing we know is that that reference population is very unlikely to be the overall adult population of Mexico. Therefore, the interpretability of the Mexican IQ scores is unclear at best.

PROBLEM THREE WITH THE MEXICAN NORMS: THE LACK OF SCORE NORMALIZATION

The lack of a meaningful reference population and the lack of a representative standardization sample are not the only limitations for the interpretation of the Mexican scores. There is no indication that the distribution of Mexican scores in the standardization sample was normal or normalized. If neither the distribution of the raw scores of the Mexican standardization sample nor that of the final scaled IQ scores in the norm tables has a normal distribution, the IQ scores are not interpretable.

As a norm-referenced score, an IQ score is interpreted as an indication of how a person performs compared to others in a particular reference population. This indication is expressed in the form of a percentile. Thus, for example, an IQ score of 70 is interpreted as performing above only 2.3% of the reference population, or being in the second percentile. In order to enable this interpretation, the exact relationship between the IQ score and the percentile must be known. For this to be known, the distribution of the IQ scores must be known. By tradition, to guarantee this known relationship, test developers have followed a practice of making sure that the scores form a normal distribution, which would provide this known relationship with percentiles. This is accomplished by either drawing a standardization sample with normally distributed raw scores or performing a series of statistical ''normalization'' transformations that would produce final scale scores that do have a normal distribution.

From page 133 of the Mexican WAIS-III technical manual, it appears that the method actually used to derive Mexican IQ scores is one that was based on the equivalence of z-scores at five anchor points between the Mexican raw scores and the scale score metric. This is a rather unusual method and the author cited ''Child (1973)'' as the source of this method but provided no reference for this citation (which we have been unable to track down). From the description, the net effect of this method is equivalent to a class of methods called ''linear standard score'' methods. Linear standard score methods produce final scores with the same shape of distribution as that of the raw scores. If the raw scores were normally distributed, the final scale scores would have a normal distribution. If the raw scores were not normally distributed, the final scale scores would not have a normal distribution. In the latter case, the corresponding percentiles for the IQ scores would be unknown.

There is no indication that the raw scores for the Mexican standardization sample formed a normal distribution, and the scaling method used as described in the manual is one that did not perform any normalization. Hence, there is no basis to convert any Mexican IQ score to a corresponding percentile; and the Mexican IQ scores are essentially not interpretable. Due to the lack of reference population information and the lack of normal distribution, we do not know what percentiles the defendant's Mexican IQ scores of 79 and 81 correspond to nor do we know the percentiles of what population.

PROBLEM FOUR WITH THE MEXICAN NORMS: LACK OF REPRESENTATION OF CERTAIN GROUPS

The developer of the Mexican norms is aware of many of the limitations of the Mexican standardization sample used to establish the Mexican norms. In the "Conclusion" section of chapter 6 in the Mexican WAIS-III technical manual, it was stated: "The same way, it is necessary to mention that the evaluator has to judge pertinence of applying the Mexican norms to the evaluated individual because in some of these cases, we address the extremes of the population, by instance people with intellectual disability, with remarkable abilities, or senior citizens. Due to the gaps in the distribution, perhaps it would be more convenient to utilize the original norms." The authors appear to indicate that "people of intellectual disability" are missing from some of the distributions in the standardization sample. Therefore, they suggest that a psychologist administering the Mexican WAIS-III test should consider using the original U.S. norms for these individuals. Since no guidance was provided as to under what circumstances people with intellectual disabilities are being or not being represented by the sample, the Mexican norms should be avoided in all situations when a person who is possibly mentally retarded is being tested. Therefore, according to the authors of the Mexican manual, the U.S. norms are more appropriate for the defendant in question. The case in which we became involved is precisely one of those situations for which the Mexican developer recommended the use of the original U.S. norms due to recognized deficiencies in the Mexican norms.

PROBLEM FIVE WITH THE MEXICAN NORMS: USE OF INCORRECT STATISTICS AND CALCULATIONS

Almost the entire technical manual for the Mexican WAIS-III IQ test is a direct verbatim Spanish translation of the U.S. technical manual for the U.S. test. The only portion of the manual that provides new technical information about the Mexican test is chapter 6, which is a relatively brief 16-page (pp. 123–138) addendum to the 229-page manual. Yet, even without consulting the translation that we obtained, we could see a number of glaring technical errors.

One example of such glaring technical errors is found in the third to the last paragraph on page 138 at the end of chapter 6. Here, it is explained that the larger SEM associated with the Mexican scores when compared against the SEM for the original U.S. IQ scores can be attributed to the smaller size of the Mexican standardization sample. This is an erroneous explanation, as there is no relationship between SEM and standardization sample size. The size of a SEM is determined exclusively by the value of the reliability coefficient and the standard deviation of the sample or population and is not related to sample size or anything else (see Equations 1 to 4). Since the standard deviations for both the U.S. and the Mexican IQ scores are set at the same value of 15, the only explanation for the larger SEM for the Mexican scores is that these scores have low reliability.

For another example, on page 125, the manual reported the following statistics: 57/43%, $\chi^2 = 4.86$; $p \geq .001$ for a comparison between the proportion of subjects in the high education-level category and that in the low education-level category. The statistics appear to have been misapplied or misreported and are not interpretable. Not all important pertinent information is reported in order to interpret the statistics. Even if the chi-square test was applied and results reported correctly, the result was misinterpreted by the author. Given that the author reported $p \geq .001$, the implicit alpha value to determine significance is .001. In such a case, the conclusion should have been that there is no or insufficient evidence of difference between the proportions of individuals with high and low levels of education. Instead, the author drew exactly the opposite conclusion—that those with a high level of education were over-represented.

There are other observable technical errors that may be as innocent as being typographical errors or may reflect more profound and consequential errors in calculation and interpretation. One such error is found in the equation on page 137. The manual provided the following equation for the calculation of scale scores:

$$\text{Equivalente escalar} = \text{DE} \times \left(\text{PE} - \frac{10}{3}\right) + \mu \qquad (5)$$

which is erroneous. The correct equation should have been:

$$\text{Equivalente escalar} = \text{DE} \times \left(\frac{\text{PE} - 10}{3}\right) + \mu \qquad (6)$$

If this is a typographical error and the correct equation had in fact been used in practice, it would be inconsequential. However, if the erroneous equation was indeed the one used in practice, the error is non-trivial. For example, for a person with a standard subscale score of say, 4, the scale IQ score would have been calculated as 110 (i.e., above average) through the erroneous equation when it should have been 70 (mentally retarded) when calculated correctly.

Without the original data, it is difficult to discern which error is a mere clerical error and which is a more profound and consequential error. However, these errors are observable even through a layer of imprecise translations by a nontechnical court interpreter. They are likely indicative of the existence of many other unobserved errors. These errors cast doubt on the correctness of the actual norming done for the Mexican scores, the correctness of the statistical information provided, and thus, the accuracy of the Mexican IQ scores obtained for the defendant.

PROBLEM SIX WITH THE MEXICAN NORMS: INAPPROPRIATE USE OF THE TRUE SCORE CONFIDENCE INTERVAL METHOD

The true score confidence interval method was used by the original WAIS-III U.S. developer. This practice was followed by the Mexican developer. This particular method is not a common method used by the overwhelming majority of tests. When used, this method will produce confidence intervals that are asymmetric around the observed test score. The net effect of this lack of symmetry is that it is in favor of a judgment that the person's true ability is closer to the average range than the extremes.

For example, take the last of the defendant's IQ scores in Table 1. His observed IQ score was reported as 81 and the 95% confidence interval as 67 to 104. This suggests that his true IQ score can be as low as 67 and as high as 104. In other words, his true IQ can be as low as 14 points below or as high as 23 points above his observed IQ of 81. As can be seen, the possible range of true IQ score on the higher end is substantially larger (23 points) than the range in the lower end (14 points). This makes it appear that the defendant's true IQ score is much more likely to be higher than 81 than it is to be lower than 81. Hence, it is in favor of a judgment of not being mentally retarded.

The alternative is the much more commonly used observed score confidence interval method, which produces symmetric confidence intervals. In the example above, had the observed score confidence interval method been used, the confidence interval of the defendant's IQ score of 81 would have been between 62 and 100. That is, his possible true IQ score would lie between either 19 points below or 19 points above his IQ of 81. Between these two methods, the former (i.e., true score confidence interval method) is not appropriate in the case of this defendant. The former method was developed by Glutting et al. (1987). On page 613 of their paper, Glutting et al. specifically addressed the issue of when to use the asymmetric true score method and when to use the symmetric observed score method. They stated that the traditional symmetric observed score confidence interval method should be used in the following situations: "Major examples include . . . use of deviation IQs in the classification of mental deficiency . . . and use of other standard scores in the identification of learning disabilities and social maladjustment."

It is clear that the traditional symmetric method that produces results suggesting a higher probability of the defendant's true IQ score being in the mentally retarded region is the correct method. The confidence intervals reported by both psychologists using either the U.S. or the Mexican norms are inappropriate and have produced a biased picture in favor of not classifying the defendant as meeting one of the MR criteria. This is a problem with both the U.S. and the Mexican norms, less so with the U.S. norms because of a much smaller confidence interval.

PROBLEM SEVEN WITH THE MEXICAN NORMS: LACK OF CONSENSUS AS TO HOW TO DEFINE MR IN MEXICO

The determination of cutoff scores for any test is inherently a matter of social policy or professional judgment. There are no inherent or natural cutoff scores that exist in some objective reality independent of human judgment. Individuals' scores on a test are on a continuum and rarely do they exist as a clearly "competent" region and a clearly "incompetent" region. Based on social policy or professional judgment, some demarcation point on the continuum is chosen as the cutoff between those who are competent and those who are incompetent. As such, cutoff scores for a particular test cannot be "discovered," "estimated," or borrowed from other tests. Instead, they must be determined through some social or professional judgmental or consensus-building process. For example, cutoff scores for high-quality professional licensure and certification exams generally employ one of a number of alternative procedures to help select appropriate cutoff scores. These procedures are designed to systematically explicate judgments from professional experts or some other authorities in numerical forms to eventually arrive at an accurate representation of their consensual judgment.

Based on the history described in a chapter by Greenspan and Switzky (2006), the process to determine the cutoff IQ score of 70–75 as the point of demarcation between mentally retarded versus not retarded was initiated by the American Association of Mental Retardation (AAMR) long before the existence of any of these formal systematic processes. However, the process followed by the AAMR was very careful, deliberative, and iterative, with authority and a high degree of professional consensus. The final cutoff score of 70–75 was arrived at after many decades of debate and revision and was clearly not taken lightly. Therefore, the cutoff criterion that individuals whose IQ scores are in the bottom 2.3% of the U.S. population are considered mentally retarded represents the best available consensual professional judgment. However, all these decades of careful deliberations, debates, revisions, and census have been about the U.S. norms only. A cutoff score of 70 determined by the AAMR means a mentally retarded individual is one who is at the bottom 2.3% of the U.S. population.

There is no evidence that the distribution of IQ scores in the adult Mexican population has ever been considered by the AAMR or by any other professional or social organization in the United States or in Mexico. Nor is there evidence of any formal or informal deliberative processes having taken place to seek consensus on what percentage of the Mexican adult population should be considered mentally retarded. Just because members of AAMR in the United States have determined after decades of deliberation that the bottom 2.3% of individuals in the United States on the U.S. IQ score distribution should be classified as mentally retarded provides no information on what percentage of the Mexican population should be classified as mentally retarded when given the Mexican IQ test. Therefore, the cutoff score of 70, which signifies "bottom 2.3%," is not applicable to the Mexican norms at all.

The fact is we have no standard in existence to determine MR on the basis of the Mexican IQ scores. Borrowing the U.S. AAMR definition of an IQ score of 70 (i.e., bottom 2.3%) and applying it to the Mexican IQ scores is convenient but unjustified and inappropriate without a process of systematic professional judgment and consensus and without considering the distribution and nature of the Mexican population. In fact, a possible powerful alternative argument may be posed as follows: Since for the same exact performance, the defendant obtained a U.S. IQ score of 70 and a Mexican IQ score of 81, this may be indicative of the equivalence of these two scores. Since the cutoff for MR on the U.S. IQ score is 70, therefore the cutoff for the Mexican IQ score for MR should be 81.

Should we apply the U.S. cutoff criterion of 70 on the U.S. norms to the defendant's Mexican IQ score to determine his MR status? This approach would mean that we believe exactly 2.3% of the population of any given nation on earth should be classified as mentally retarded. Should we instead consider a Mexican IQ score of 81 as the cutoff for MR since it corresponds to a U.S. IQ score of 70? This approach would mean that we believe any human being performing at a level equivalent to the bottom 2.3% of the U.S. population should be classified as mentally retarded. Without any formal social or professional dialogue, debate, deliberation, or consensus, there is no way to determine which of these or some other criterion is appropriate.

Therefore, even if the Mexican IQ scores were reliable, free from errors, with no sampling or norming problems, we would still be left with a problem of having no basis to make a judgment about MR due to a lack of any standard for MR with the Mexican IQ scores.

CONCLUSION

When considering all of the many problems that we have identified with the Mexican WAIS-III norms, one is led to the inescapable conclusion that the IQ scores obtained for the defendant based on the Mexican norms are likely to be unreliable and not interpretable. One can also conclude that there is not a meaningful or authoritative cutoff score for the Mexican IQ scores that can be used to determine the defendant's MR status. Compounding these factors was the potential presence of many unknown technical errors. It is quite conclusive that the use of the Mexican norms is inappropriate, and the results are likely to be invalid.

Throughout the above discussion, we have focused on inadequacies or errors in score reliability, sampling, norms, cutoff standards, scaling, and other technical problems surrounding the scores and norms. We made no attempt to examine issues related to adequacy of evidence for validity. The general assumption implicit in the Mexican manual which is tacitly accepted by all involved in the case is that, since the Mexican version of WAIS-III is a verbatim translation of the original U.S. WAIS-III, with a few minor adjustments to account for cultural differences, the Mexican WAIS-III is a valid measure of intelligence. It is further assumed that the construct of intelligence reflected by the Mexican WAIS-III has the same internal structure, same nomological net, same dimensionalities, and same meaning and function as the construct of intelligence reflected by the U.S. WAIS-III scores. These assumptions have been taken on face value and have never been tested.

There is in fact no evidence to support the assumption of construct equivalence between the two versions of the IQ test. No back translation was performed to ensure semantic equivalence, and no formal independent validation study has been reported for the Mexican version. These issues of validity were never raised in the case since the technical deficiencies surrounding the scores and norms for the purpose of the hearing were already overwhelming. However, should the problems of measurement and interpretability raised in this paper be resolved in the future, there will need to be additional investigations to gather either independent evidence of validity for the Mexican WAIS-III or evidence of construct equivalence between the U.S. and the Mexican versions of the WAIS-III or both.

We are aware of several other Atkins proceedings around the United States in which the Mexican WAIS-III norms have been used, sometimes even where the defendant is from a country other than Mexico (e.g., Colombia). In all of those cases, the use of the Mexican norms has been justified in knee-jerk form by Spanish-speaking psychologists on the grounds of cultural sensitivity. This argument is perverse, however, if the norms that are used are worthless. All psychologists have an absolute obligation to ensure that their diagnostic conclusions are based on valid and reliable information from psychological tests. To ignore this issue is to mark oneself as an incompetent practitioner. Basic ethics also requires the publishers of the Mexican WAIS-III to withdraw from the market a set of test norms that have the potential to do much harm, especially in death penalty cases.

ACKNOWLEDGMENT

We are grateful to Professor Marley Watkins of Arizona State University for his input regarding the prevalence of the use of "true score confidence interval" among practitioners as well as general practices of translating IQ tests. Special thanks are expressed to attorney Jay Grant, who provided us with the resources needed to conduct this analysis and for recognizing the importance of getting this information to a wider public.

REFERENCES

Glutting, J. J., McDermott, P. A., & Stanley, J. C. (1987). Resolving differences among methods of establishing confidence limits for test scores. *Educational and Psychological Measurement*, *47*, 607–614.

Greenspan, S., & Switzky, H. N. (2006). Forty-four years of AAMR manuals. In H. N. Switzky & S. Greenspan (Eds.), *What is mental retardation? Ideas for an evolving disability category in the 21st century* (pp. 3–28). Washington, DC: American Association on Mental Retardation.

Heaton, R. K., Taylor, M., & Manley, J. (2003). Demographic effects and use of demographically corrected norms with the WAIS-III and WMS-III. In D. S. Tulsky, R. K. Heaton, G. J. Chelune, R. J. Ivnik, R. Bornstein, A. Prifitera, D. H. Saklofske, & M. F. Ledbetter (Eds.), *Clinical interpretation of the WAIS-III and WMS-III* (pp. 181–210). San Diego, CA: Academic Press.

Mercer, J. (1988). Ethnic differences in IQ scores: What do they mean. *Hispanic Journal of Behavioral Sciences*, *10*(3), 199–218.

The Psychological Corporation. (1997). *WAIS-III, WMS-III technical manual*. San Antonio, TX: Author.

Tulsky, D. & Zhu, J. (2003). *Escala Wechsler de Inteligencia para Adultos-III*. Moderno: Mexico DF (Selected chapters translated by T. Rosado, court translator).