2013

IMPACT OF RACE ON JUVENILE FORENSIC ASSESSMENT: CLINICIAN PERCEPTIONS OF ADEQUATE TEST NORMS

Nathan E. Cook
University of Rhode Island, nathanecook@gmail.com

IMPACT OF RACE ON JUVENILE FORENSIC ASSESSMENT:

CLINICIAN PERCEPTIONS OF ADEQUATE TEST NORMS

BY

NATHAN E. COOK

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

2013

DOCTOR OF PHILOSOPHY DISSERTATION

OF

NATHAN E. COOK

APPROVED:

Thesis Committee:

Major Professor David Faust

Leslie Mahler

Jasmine Mena

Lisa Weyandt

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2013

ABSTRACT

The psychological assessment of racial and ethnic minority groups is often substantially limited by the lack of adequate normative data for these groups. This study examined the impact that race has on forensic psychologists' (*N*=145) diagnostic decision making as well as judgments of the quality of normative data. It was hypothesized that the forensic psychologists would accept lower quality normative data for African American youth compared to White youth. However, although the quality of the test norms influenced the dependent measures in the expected direction, no significant interaction was noted between norm quality and youth's race. Participants judged the likelihood of disorder and quality of norms similarly for White and African American youth, and also expressed similar levels of confidence in their diagnostic judgments regardless of youth's race. These findings are encouraging in that they suggest clinicians did not apply differential standards when appraising test norms for African American youth compared to White youth. Clinical implications and future research directions are discussed.

ACKNOWLEDGMENTS

I would like to sincerely thank my major professor Dr. David Faust, not only for his assistance with this project but, more generally, for his interest in and support of my intellectual growth. I am incredibly fortunate to have profited from his mentorship and guidance. Also, I would like to extend my gratitude to my core committee members, Dr. Leslie Mahler, Dr. Jasmine Mena, and Dr. Lisa Weyandt, for their support of this project as well as other milestones throughout my graduate training. Finally, appreciation is owed to the additional members of my defense committee, Dr. Jill Doerner (defense chair) and Dr. Jerry Cohen.

DEDICATION

This dissertation is dedicated to my wife. Thank you for your unconditional love and unfailing support, which made this dissertation as well as other academic and professional endeavors possible. And to my daughter, whose sense of curiosity and wonder inspires my own. Thank you for providing levity and joy in times of stress, helping me to keep things in perspective as I pursue my academic goals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The first specialized juvenile court was opened in Cook County, Illinois in 1899. Grisso (1998) noted that within 10 years a clinical evaluation service was established by two professionals, a neurologist and psychologist. (Interestingly, the first text providing a systematic, conceptual approach to such evaluations, i.e., Grisso's 1998 *Forensic Evaluation of Juveniles*, was not published until nearly a century later.) These early evaluations employed a comprehensive, multidisciplinary approach that aligned with the early philosophical foundations of the juvenile court, namely the doctrine of *parens patriae*, a beneficent and rehabilitative approach to delinquent youth (Grisso, 1998). Since its inception the ethos of the juvenile court has undergone major shifts (see Grisso, 1996). Still, despite changes in the types and function of assessments provided to the juvenile court, mental health professionals remain involved in juvenile court proceedings (Otto, Borum, & Epstein, 2012).

Although legal and philosophical shifts in juvenile justice policy have produced more specialized forms of evaluation (e.g., competency to waive rights, fitness to stand trial), an important role of psychologists is to assess youth for mental health problems and therapeutic intervention using methods supported by research (Hoge, 2008; Vincent, Chapman, & Cook, 2011). Further, in such evaluations psychologists report frequently using norm-referenced psychological tests (Archer, Buffington-Vollum, Stredny, & Handel, 2006; Viljoen, McLachlan, & Vincent, 2010), or

measures for which one or more reference groups are used to compare an individual's score to that of others in order to draw inferences (AERA, APA, & NCME, 1999).

CHAPTER 2

REVIEW OF LITERATURE

Grisso (2005a) noted that juvenile justice authorities have important custodial, due process, and public safety obligations to identify mental disorders within delinquent youth. Although the most pressing issue facing juvenile justice administrators 15 years ago was public safety, presently it is mental health (Grisso, 2005b). This finding is likely due to recent evidence of alarmingly high rates of mental health problems among juvenile justice involved youth (see Teplin, Abram, McClelland, Dulcan, & Mericle, 2002; Abram, Teplin, & McClelland, 2003). Teplin and colleagues (2002) found that, among juvenile detainees, the most common maladies were substance use disorders (47% and 51% for females and males respectively), followed by disruptive behavior disorders (i.e., oppositional defiant disorder and conduct disorder; 31% and 41%), and then anxiety disorders (21% for both sexes). Attention-deficit/hyperactivity disorder was present in 21.4% of female and 16.6% of male detainees (Teplin et al., 2002). Grisso's (2005b) literature review indicated that rates of mental disorders in the juvenile justice system are as high as 60% or 70%, which is two to three times greater than prevalence rates among general population youth.

Research evidence is also beginning to accumulate that suggests possible racial and gender differences in mental health problems amongst juvenile justice populations. Specifically, female youth in the juvenile justice system exhibit greater mental health problems and more severe pathology than males (Abram et al., 2003;

Kataoka et al., 2001; Vincent, Grisso, Terry, & Banks, 2008). For example, Abram and colleagues (2003) reported that in a juvenile detention facility significantly more females (57%) than males (46%) met criteria for two or more disorders including substance abuse and conduct disorder. The research concerning racial/ethnic differences has been mixed. Some authors have reported higher rates among White youth in comparison to African American youth (Teplin et al., 2002) while others reported the opposite outcome (Rawal, Romansky, Jenuwine, & Lyons, 2004). Still others reported that rates of mental disorder are similar for White and African American youth (Shufelt & Cocozza, 2006). It is consistently reported, however, that White youth are more likely to be referred for mental health services than minority youth (Hawkins, Laub, Lauritsen, & Cothern, 2000; Smedley, Stith, & Nelson, 2003; Teplin et al., 2005).

In addition to mental health problems, there is mounting evidence that juvenile justice involved youth differ from general population youth in other important ways. For example, youth that are arrested tend to have lower school achievement (Geib et al., 2011). It is unclear, however, if this finding is confounded since mental health problems also contribute to underachievement in school (Patel, Flisher, Hetrick, & McGorry, 2007). Further, in a large sample of youth, lower income was associated with an increased likelihood of juvenile delinquency (Bright & Jonson-Reid, 2008). Although the great many studies comparing juvenile court involved youth and community samples use groups matched for sociodemographic variables such as age or education, thus obscuring potentially important demographic differences, there is considerable evidence that court-involved youth differ in important ways from

community youth. Such differences could limit the generalizability of psychological tests developed and normed for use with non-juvenile justice involved youth if true ability and performance on the test is related to sociodemographic variables such as SES or education.

Given the clear need for mental health evaluations of justice-involved youth and the sociodemographic differences found in this population, juvenile forensic assessment experts have stressed the importance of critically appraising and applying appropriate test norms (Grisso, 2005a). Psychological test norms are the standard for comparison when interpreting an individual's score on a psychological test against others, usually a group of healthy or presumably normal individuals (Mitrushina, Boone, Razani, & D'Elia, 2005). Selecting adequate, representative test norms is critical to optimize accuracy when interpreting psychological test results and allows evaluators to judge what represents "normal" performance on a measure (AERA, APA, & NCME, 1999; Mitrushina et al., 2005). Unfortunately, the professional and financial support necessary to conduct high quality normative research have often been lacking (Mitrushina et al., 2005).

Use of weak or inadequate normative data can have considerable, negative effects on psychological testing and interpretation. Norms based on small samples, obsolete norms, or an inappropriate demographic match between the client and the normative sample will often reduce accuracy rates and thus create multiple negative consequences. For example, the wrong condition may be identified leading to the wrong treatment, or a normal individual may be misidentified as abnormal leading to unnecessary treatment that may carry risks. Inappropriate norms can both

overpathologize, i.e., make normal people appear impaired, disturbed, or worse off than is actually the case, and can also underpathologize a given youth which would present different risks (see Faust, Ziskin, & Hiers, 1991). Additionally, the score distributions across normative sub-groups (e.g., based on sociodemographic dimensions such as education, age, culture, and ethnicity) can differ (Wood, Garb, & Nezworski, 2007). Frequently, individuals might score in the impaired range when one set of norms is used while scoring in the normal range given a different set of norms (Mitrushina et al., 2005).

Differences in scores based on sociodemographic characteristics raise concern about the potential selection of inappropriate norms, leading to increased error (Kalechstein, van Gorp, & Rapport, 1998; Mitrushina et al., 2005). In many cases the differences in test score interpretation based purely on the normative sample used can be large. For example, Kalechstein and colleagues (1998) reported substantial variability in the standard scores derived from neuropsychological tests based on the norm sample used. For one test the same raw score was classified as Average, Low-Average, and Impaired across various norm samples. Of note, variability was associated with sociodemographic characteristics such as gender, age, and education. That is, greater variability in test score interpretation was found for men, younger individuals, and for individuals with fewer years of education. The authors note that it is unclear whether the variability is due to "true" differences or methodological artifacts such as measurement error or sampling practices including use of exclusionary criteria, i.e., factors that screen out or exclude certain subjects in normative data sets, such as medical history or current health status (Kalechstein, van

Gorp, & Rapport, 1998). What does appear clear is that failure to account for sociodemographic variables in normative interpretation can lead to inconsistent, potentially opposing conclusions.

Given the potential consequences of using inappropriate norms, professional organizations have addressed their importance in clinical practice. The *Standards for Educational and Psychological Testing* (*Standards*) devote Section 4, "Scales, Norms, and Score Comparability" to the topic (AERA, APA, & NCME, 1999). The *Standards* stress that the reference group used in the normative sample "be carefully and clearly described" (p. 51), thus allowing evaluators "to judge the appropriateness of the norms" (AERA, APA, & NCME, 1999, p. 55). The American Psychological Association's (APA; 2002) *Ethical Principles of Psychologists and Code of Conduct* do not make specific mention of "norms" but devote Standard 9 to the issue of psychological assessment. Within this section psychologists are encouraged to "use assessment instruments whose validity and reliability have been established for use with members of the population tested" (p. 1071).

Echoing the considerations raised by the *Standards* and *Ethics Code*, several authors have discussed the importance of sociodemographic match between the normative sample and the client. Regarding neuropsychological assessment, Mitrushina and colleagues (2005) noted the importance of demographic similarity. In the context of juvenile forensic assessment both Grisso (2005a) and Koocher (2006) mentioned the importance of appraising demographic match in test norms. Grisso (2005a) further observed that "the current state of the art provides almost no guidance regarding what you should look for in a tool when considering its use with youths

from various ethnic groups" (p. 82). An important caveat regarding demographic match is the potential for test bias to influence outcomes when "true" standing or ability is differentially related to test scores for different groups. In certain cases a given test score might be related to demographic characteristics of the examinee but be unrelated to actual ability. If this is the case even including substantial proportions of demographically similar individuals in the normative samples will not alleviate the test bias.

Even assuming a test is minimally influenced by bias, little scientifically grounded guidance is available regarding how evaluators should determine if norms are appropriate or if one set of norms is more appropriate than another set. Further, there are typically multiple indicators or dimensions that one might consider when appraising normative data sets. Often these dimensions have uncertain or unknown levels of impact on test accuracy, their degree of overlap or redundancy may be unclear, and they may conflict with one another. Thus, psychologists are presented with a complex decision task that must be performed in the absence of adequate scientific knowledge. Generally the human mind does not perform well when integrating complex information (Faust, 2008), even when it is relatively complete or unambiguous.

Mitrushina and colleagues (2005) described four dimensions meriting consideration when appraising and selecting test norms: a) sample size, b) similarity between normative study and evaluator in terms of method of administering and scoring the test, c) how recently the normative data were gathered, and d) the similarity between demographic characteristics of the norm sample and the client. In

many circumstances, when multiple normative datasets are available, standing on these factors will conflict and/or be uncertain. For example, one set of norms may be superior to another on dimension A, inferior on B, and of unclear relative standing on C and D. Unfortunately, there is almost no scientific guidance available for resolving such conflicts or ambiguities, such as how heavily to weigh one dimension versus another, or whether differential weighting should be used at all.

One suggested approach to norm selection involves determining which normative data to use prior to testing an individual (Kalechstein, van Gorp, & Rapport, 1998). This approach seeks to remove any temptation to choose norms that will confirm one's subjective diagnostic impressions (Kalechstein, van Gorp, & Rapport, 1998). However, clinicians are afforded little specific, scientific guidance regarding how to judge the appropriateness of test norms along objective dimensions. There are limited *explicit* guidelines available (e.g., "Match norms with the examinee's age as closely as possible") and perhaps more importantly, no *properly founded* guidelines for evaluators to follow when appraising how well test norms match a given client or how to select among competing normative data (see Faust, 2005), with the partial exception being sample size (see Bridges & Holler, 2007).

Without formal guidance, evaluators likely rely on their clinical judgment or impressionistic methods (Kalechstein, van Gorp, & Rapport, 1998). With this approach to decision making psychologists rely on subjective weighting or inference when combining data (Meehl, 1954). When employing impressionistic strategies, clinicians process and combine data in their head, using their own judgment rather than scientifically validated procedures. Echemendia and Harris (2004) reported a

striking example. They surveyed practicing clinical neuropsychologists and reported that 57% did not compare the performance of monolingual Spanish speakers to norms at all but used their own clinical judgment instead. Contrary to such questionable practices, for over 50 years authors have described the limitations of using subjective clinical judgment in decision making, including decreased predicative accuracy compared to alternative approaches (Dawes, Faust, Meehl, 1989; Faust, 1984; Meehl, 1954).

In addition to decreased predictive accuracy, reliance on clinical judgment may provide an open field for biases to influence the evaluation and selection of test norms for minority group clients. Researchers on social judgment have proposed the *shifting standards* model (Biernat & Manis, 1994); this model posits that psychologists "may use unacknowledged cognitive schemas as reference points when making subjective judgments about members of stereotyped social groups" (Gushue, 2004, p.398). An important implication of this model is that the reference point shifts for different groups. For example, a study of graduate students in counseling and clinical psychology found that given identical sympotmotology, hypothetical African American clients were reported to be less symptomatic compared to White clients (Gushue, 2004). These results suggest that mental health professionals may adjust their standards for appraising symptom severity based on a client's race. Pertinent to the present study, such research suggests that a similar process may lead psychological evaluators to apply shifting standards to the appraisal of test norms and increase their willingness to accept lower quality normative data for social groups that are negatively stereotyped. Specifically, the reference point or criterion for "good" normative data

may in fact be different for majority compared to minority clients. This is particularly relevant for members of minority groups since normative data for such groups may be totally nonexistent or severely limited by, for example, exceedingly small sample sizes of just four or five individuals (Faust, Ahern, & Bridges, 2012).

Concerns regarding racial/ethnic bias in clinical and diagnostic judgment are especially salient for juvenile forensic assessment given consistent evidence of "disproportionate minority contact," which refers to the finding that minority youth are overrepresented at each level of the juvenile justice system (see Kempf-Leonard, 2007; Piquero, 2008). Data from 2002 indicate that while Black youth accounted for 16% of the general population, they accounted for 29% of youth adjudicated in juvenile court, 33% of youth placed out of home, and 35% of youth waived to adult criminal court (Kempf-Leonard, 2007). More recent data indicate that although White youth represent the largest percentage of the population they are detained at lower rates than Black and Latino youth (National Council on Crime and Delinquency, 2007).

In addition to disparities in overall involvement with the justice system, there appear to be racial and ethnic disparities regarding mental health referral and utilization among juvenile justice involved youth. As previously noted, research suggests that White youth are more likely to be referred for mental health services than minority youth (Hawkins, Laub, Lauritsen, & Cothern, 2000; Smedley, Stith, & Nelson, 2003; Teplin et al., 2005). Further, research also suggests that White delinquent youth are more likely to receive psychiatric and medical services than Black youth (Otnow Lewis, Balls, & Shanok, 1979). More recently, a study from the

Illinois juvenile justice system reported that White youth had higher rates of past, current, and overall mental health service utilization compared to Black and Hispanic youth (Rawal et al., 2004).

In summary, there is currently no consensus regarding what constitutes adequate norms for psychological assessment, and this is especially concerning for the assessment of under-represented and under-served groups. As noted above, normative data for minority groups are often absent or severely limited (Faust, Ahern, & Bridges, 2012). In addition, there is limited knowledge regarding if and how juvenile forensic evaluators judge the quality or appropriateness of available test norms. To address these issues, the present study examined whether practicing juvenile forensic evaluators apply higher standards to test norms when evaluating White youth compared to minority youth. The primary study hypothesis is that decreases in diagnostic confidence and ratings of norm quality will interact with the youth's race such that clinicians will rate poorer norms as being of higher quality for African American youth compared to White youth and will express higher diagnostic confidence when presented with poorer norms for African American youth compared to White youth. A secondary hypothesis is that clinicians will estimate higher likelihood of disorder, report more confidence, and rate norm quality higher when presented with psychological test data and as the quality of test norms provided increases. Lastly, it is hypothesized that clinicians will estimate lower likelihood of disorder for African American youth compared to White youth, given previous research suggesting differential standards when judging psychopathology for White and African American clients (Gushue, 2004).

CHAPTER 3

METHODOLOGY

Participants

Forensic psychologists who conduct mental health assessments of juvenile justice youth were recruited via email solicitation. Invitations to participate were distributed to the members of Division 41 of the American Psychological Association (American Psychology-Law Society) and subscribers of both the Division 12 (Clinical Psychology) and PSYLAW (a psychology and law discussion group) listservs. The solicitation asked that recipients forward the survey link to other professionals and colleagues. To be eligible for participation respondents had to be licensed clinicians and conduct mental health assessments of juvenile justice involved youth at least twice per month. There are certain advantages in using modern technology such as email to disseminate a survey. A major drawback is the challenge of determining or estimating response rate. Unfortunately, for this study the number of eligible individuals who did and did not respond is unknown. This raises some concern regarding the representativeness of the sample. Still, given the preliminary nature of this study, this recruitment strategy was considered the best practical compromise.

Given results from a previous study with a similar target population (Archer et al., 2006) the present study was expected to elicit a demographically similar sample. Specifically, respondents to Archer and colleague's (2006) study were mostly male (approximately 60%) and White (91%). Thus, a statement was in the solicitation

email attempting to encourage women and racial/ethnic minority evaluators to participate.

A total of 145 respondents completed the survey. Demographic characteristics of the sample are presented in Table 1. Respondents were predominantly male (59.3%) and overwhelmingly self-identified as White (91.7%). Respondents varied in terms of professional and forensic assessment experience. The mean time since participants earned their degree was 17.3 years (*SD*=12.0, range 0 to 40 years) and had 16 years of forensic assessment experience (*SD*=10.7, range 0 to 40 years). The median number of assessment conducted in the past year[1] was 49.5, of which an average of 48.4% (*SD*=27.9) involved a minority evaluee.

Procedure

This study employed a 2 (youth race) by 4 (test presence/norm quality) between subjects design. Participants who met the eligibility criteria were randomly assigned to one of the eight conditions. *A priori* power analysis suggested a target cell size of 30 participants per condition. Obtained cell sizes ranged from 13 to 23.

Each condition included a brief case vignette which described a hypothetical youth that participants were evaluating for the court (see Appendix A). The vignettes include only a hypothetical female youth in order to control for the potential influence of youth gender and because of research literature suggesting that females in the juvenile justice system tend to have higher rates and more severe mental health problems compared to males (Abram et al., 2003; Kataoka et al., 2001; Vincent et al., 2008). The female youth in the scenario was described as undergoing a psychological

[1] The number of assessments conducted in the past year was skewed so the median provides the most appropriate measure of central tendency.

evaluation for a potential substance use disorder. A psychodiagnostic task was chosen given evidence regarding racial biases in diagnosis that involve both over- and underpathologizing minority group members (see Whaley & Geller, 2007 for a review). Substance use disorder was chosen because in a large, national sample of juvenile justice involved youth, racial differences were found in the percentage of youth reporting problematic alcohol and drug use (Vincent et al., 2008). That study found White youth were most likely to report substance use problems (32%), followed by Hispanic youth (27%) and African American youth (19%). Differences in base rates of a disorder impact test norm cutoff scores such that the less common the disorder in a given subgroup (and all else being equal) the higher the cutoff score should be set, and failure to account for this difference leads to decreased diagnostic accuracy (Meehl & Rosen, 1955). Also, race has not been found to influence treatment recommendations for alcohol or drug use amongst juvenile court involved youth (Breda, 2003). Thus, this hypothetical case was one in which race itself does not seem to sway juvenile court decision makers but presents a mental health concern where racial disparities have been reported and likely bear on the composition of test norms and use of clinical cutoff scores. As a result, application of normative data that do not provide a good sociodemographic match for a given client would likely lead to diminished interpretive accuracy.

The vignettes described the demographic characteristics of the youth identically with the exception of the first independent variable, youth's race. This variable had two levels. Participants were assigned to a condition in which the youth was either

described as "White" or "African American." Additional information about the youth, such as age and sex, was held constant across conditions.

The second independent variable was the presence of a psychological measure and the quality of the measure's norms. This variable had four levels. For one condition (*No Test)* participants received only the base rate of the potential disorder (50%) with no additional information. Participants in the three remaining conditions were told that a psychometric instrument was available to help them address the referral question in the case. These conditions included manipulations of certain characteristics of the instrument's normative data. The normative dimensions presented to participants were selected given that published texts on test norms highlight these dimensions as important to consider when evaluating norms (e.g., Mitrushina et al., 2005). Specifically, three aspects of norms were varied in this study: sample size, obsolescence, and age match.

Bridges and Holler (2007) reported that a sample size of at least 50 is necessary to render confidence intervals around test scores that are clinically useful. In the *Ambiguous Norms* condition the normative data had a subgroup sample size of 35 youth that match the hypothetical youth's racial background. As the minimum adequate sample size suggested by Bridges and Holler (2007) is 50, a sample size of 35 is intended to be "ambiguous" and thus produce a situation of uncertainty regarding how adequate the data are based on this dimension. The *Small Sample Size* condition was identical to the *Ambiguous Norms* condition except that the size of the normative database was only15 subjects.

In regards to obsolescence, or recency of data, a widely used text (Strauss, Sherman, & Spreen, 2006) suggested that normative data become obsolete in about 15 to 20 years. The age of the norms was presented as 6 years old in the *Ambiguous* and *Small Sample Size* conditions. In the final condition (*Poor Norms*) the age of the data was 26 years. Lastly, the age range of the subjects in the normative group included the hypothetical youth's age in the *Ambiguous* and *Small Sample Size* conditions but did not include subjects that matched the hypothetical youth's age in the *Poor Norms* condition. Thus, the normative conditions were roughly ordered in quality from *Ambiguous*, which, in relative terms, was intended to have the most favorable characteristics, to *Small Sample Size*, to *Poor Norms*, which was intended to have the least favorable characteristics (i.e., small sample size, obsolete data, and poor age match).

Participants were further provided with the normative sample's mean score (100), standard deviation (15), and two cut-off scores ("120 to 129 - Borderline range; above 130 - Clinical range"), as well as the score of the youth being evaluated (135). On the measure, higher scores represented greater pathology. The youth's score was presented as more than 2 standard deviations above the norm sample mean, a score commonly considered "clinically significant" and also within the range of error that could occur based purely on norm selection (see Kalechstein et al., 1988). This information was held constant across the three norm conditions.

Participants then completed a brief questionnaire. First, participants made a diagnostic judgment regarding the likelihood of the hypothetical youth having a substance use disorder, rated as a probability estimate between 0 and 100%.

Participants then rated their level of confidence in the diagnostic judgment on a 7-point scale ranging from 1 (Not At All Confident) to 7 (Extremely Confident). Next, in all but the *No Test* condition, participants rated the quality of the normative data on a 7-point scale ranging from 1 (Very Poor) to 7 (Very Good). These three ratings represented the primary dependent variables.

Participants were also asked to provide demographic information including sex, race, ethnicity, years of experience conducting general psychological assessment, years of experience conducing forensic assessment, number of assessments conducted in the past year, percent of assessments in the past year involving a minority evaluee, and amount of professional training experience focused on multicultural issues. Lastly, participants were asked to self-appraise their multicultural competence on a scale from 1 (Not At All Competent) to 7 (Extremely Competent). The eight conditions did not differ significantly on any of these demographic characteristics.

Analyses

To test the main hypothesis, two-way analyses of variance (ANOVA) were conducted examining the effect of the youth's race (White or Black) and the presence/quality of normative data (*No Test*, *Ambiguous Norms*, *Small Sample Size*, or *Poor Norms*). Mean differences in perceived likelihood of disorder, confidence in that diagnostic likelihood, and quality ratings of the normative data were examined. This analytic approach allowed for examination of both main effects for each IV as well as the interaction between IVs. If statistically significant differences were found Tukey post-hoc comparisons were conducted to see which groups differed and determine the

direction of the difference. Effect sizes were also computed to characterize the magnitude of group differences.

Additional exploratory analyses were conducted to examine potential correlations between and among respondent experience and training variables (e.g., years of experience and self-appraised multicultural competence) and the primary dependent measures. Lastly, mean differences between respondent sex and self-identified racial/ethnic groups on experience and training variables were examined. Some demographic variables were not normally distributed (e.g., hours of didactic training focused on multicultural issues) and thus, nonparametric methods were used when appropriate. Normality was assessed by conducting Kolmogorov-Smirnov Tests of Normality, examining skewness and kurtosis values, and by visually inspecting histograms.

CHAPTER 4

FINDINGS

Descriptive statistics for the full factorial design are presented in Table 2. Source tables for the three ANOVAs testing primary study hypotheses are provided in Table 3.

Likelihood of Substance Use Disorder

Participants' estimates of the likelihood that the youth had a substance use disorder were examined with a two-way ANOVA (see Figure 1). The analysis yielded a significant main effect for test presence/quality, F (3,137) = 14.00, $p < .001$ (see Table 4). Post-hoc Tukey tests indicated that participants in the *No Test* condition reported a significantly lower estimate of likelihood compared to all other conditions. Effect sizes (Cohen's *d*) for the pairwise comparisons between the *No Test* condition with the three other conditions (*Poor Norms*, *Small Sample Size*, and *Ambiguous Norms)* ranged from medium to large (d = -0.71, -1.05, and -1.75 respectively). In addition, participants in the *Poor Norms* condition reported significantly lower estimates of likelihood compared to the *Ambiguous Norms* condition (d = -0.75). This difference represented a medium effect size. The main effect for youth's race and the interaction effect were not statistically significant nor was there a trend to suggest that youth's race influenced ratings of likelihood.

Confidence in Diagnostic Judgment

Participants' confidence in the judgment of disorder likelihood was also examined with a two-way ANOVA (see Figure 2). This analysis similarly yielded a main effect

for test presence/norm quality, F (3, 137) = 12.66, $p < .001$ (see Table 4). Follow-up procedures showed that participants in the *No Test* condition reported significantly less confidence in their judgment than both the *Ambiguous Norms* (d = -1.29) and *Small Sample Size* (d = -0.98) conditions, both large effect sizes. Additionally, participants in the *Poor Norms* condition reported significantly less confidence than the *Ambiguous Norms* condition (d = -1.03), representing a large effect size. The main effect for youth's race and the interaction effect were again non-significant with no trend to suggest that youth's race influenced ratings of confidence.

Appraisal of Norm Quality

Participants' appraisal of the quality of the test norms was examined with a two-way ANOVA (see Figure 3) and again yielded a main effect for test presence/norm quality, F (2, 98) = 16.49, $p < .001$ (see Table 4). Post-hoc procedures indicated that the *Ambiguous Norms* were rated as significantly higher in quality than both the *Small Sample Norms* (d = 0.83) and the *Poor Norms* (d = 1.47). Both of these differences represented large effect sizes. As with the other two analyses, the main effect for youth's race and the interaction effect were non-significant with no trend to suggest youth's race influenced ratings of norm quality.

Exploratory Analyses

Several exploratory analyses were conducted to examine for associations between and among participants' experience and training variables as well as the three dependent measures. First, a correlation matrix was computed. The experience and training variables were not associated with any of the three main dependent variables. For example, years of experience as a forensic evaluator did not appear related to

ratings of confidence or norm quality. However, some interesting associations did emerge among the training and experience variables and are summarized in Table 5. Years since graduation, years of overall assessment experience, and years of forensic assessment experience were significantly negatively correlated with the amount of didactic and supervision experience focused on multicultural issues (ρ ranged from -.24 to -.33) such that more experienced clinicians reported less training experiences focused on multicultural issues. The amount of didactic (ρ =.27, p = .002) and supervision (ρ =.36, p < .001) training focused on multicultural issues was positively correlated with the percentage of assessments conducted in the previous year involving minority clients. Clinicians with more training focused on multiculturalism appeared more likely to evaluate minority clients. Finally, self-appraised multicultural competence was positively correlated with percentage of assessments conducted in the previous year involving minority clients (ρ = .28, p = .001), hours of didactic multicultural training (ρ = .28, p = .001), and hours of supervision focused on multicultural issues (ρ = .23, p = .009). These findings suggest that clinicians conducting a higher percentage of evaluations with minority clients and with more hours of training focused on multicultural issues rate themselves as more multiculturally competent, though it should be noted that the magnitude of these associations are small using Cohen's (1988) interpretive guidelines.

Next, mean differences on the training and experience variables between participant sex and self-identified race/ethnicity were examined. There were no group differences on self-appraised multicultural competence between racial/ethnic groups or between sexes. Statistically significant mean differences were found between sexes

and racial/ethnic groups on prior training focused on multicultural issues. Specifically, female participants (*Mdn* = 75) reported significantly more hours of didactic training focused on multicultural issues than male participants (*Mdn* = 62.5), two-tailed Mann-Whitney *U* = 1,535.50, *Z* = -2.94, *p* = .022. Female participants (*Mdn* = 49) also reported significantly more hours of supervision focused on multicultural issues than male participants (*Mdn* = 24), two-tailed Mann-Whitney *U* = 1,399, *Z* = -2.65, *p* = .008. Of note, female participants (*M* = 12.27, *SD* = 11.23) graduated significantly more recently compared to male participants (*M* = 20.71, *SD* = 11.41), *t* (132) = -4.24, *p* <.001.

A Kruskal-Wallis test revealed significant differences between racial and ethnic groups on hours of didactic multicultural training, *H* (4) = 12.09, *p* = .017. Follow-up Mann-Whitney *U* tests revealed that participants who self-identified as Hispanic (*Mdn* = 200) reported more hours of didactic training than White participants (*Mdn* = 73), *U* = 13.5, *Z* = -2.73, *p* = .001. Lastly, the differences between racial/ethnic groups on hours of supervision focused on multicultural issues approached significance, *H*(4) = 9.03, *p* = .06. Follow-up Mann-Whitney U tests revealed that, as with didactic training, Hispanic participants (*Mdn* = 173) reported significantly more hours of clinical supervision focused on multicultural issues than White participants (*Mdn* = 25), *U* = 20.5, *Z* = -2.6, *p* = .003.

CHAPTER 5

CONCLUSION

Consistent with a subset of the research hypotheses, the presence of a psychological test and quality of the tests norms influenced the dependent measures in the expected direction, with large effect sizes observed. Specifically, participants judged a higher likelihood of disorder and more confidence in the presence of a test, as well as with higher quality test norms (i.e., from *Poor Norms* to *Ambiguous Norms*). Also, as the characteristics of the presented norms improved, participant ratings of quality increased accordingly (i.e., lowest ratings of quality for *Poor Norms* and highest ratings for *Ambiguous Norms*). These findings suggest that clinicians attended to the psychological test data and quality of test norms and adjusted their judgment practices and confidence accordingly.

Contrary to the hypothesized outcome, no significant interaction was noted between test presence/norm quality and youth's race. Participants judged the likelihood of disorder and quality of norms similarly for White and African American youth, and also expressed similar levels of confidence in their diagnostic judgments regardless of youth's race. These findings are encouraging. They suggest clinicians did not apply differential standards to test data for African American youth compared to White youth, such as exhibiting more leniency toward psychometric weaknesses or limitations with minority groups. Further, the hypothesis that youth race would influence psychologist judgments regarding likelihood of disorder was not supported.

Participants' ratings of the likelihood of disorder did not differ for White or African American youth. This finding differs from previous research suggesting that clinicians overpathologize African American clients (Garb, 1997; Gushue, 2004).

Exploratory analyses suggested that more experienced clinicians reported fewer hours of didactic training and supervised clinical experiences focused specifically on multicultural issues, likely reflecting growing emphasis on cross-cultural training in recent years. Such training appears particularly important for forensic evaluators given that cultural issues relevant to assessment are not widely covered in forensic psychology textbooks (Powell & Bartholomew, 2003) nor is race or ethnicity addressed in much of the published research in prominent forensic psychology journals (Carter & Forsyth, 2007). Also, clinicians with more hours of didactic and supervision training reported evaluating a higher percentage of minority clients and rated themselves as more multiculturally competent. However, increased self-appraised multicultural competence, greater experience evaluating minority clients, and more hours of specialized multicultural training were not related to participants' diagnostic judgments and ratings of norm quality. Additionally, female clinicians reported significantly more didactic and supervised clinical training focused on multicultural issues than male clinicians. Finally, Hispanic clinicians reported more didactic and supervision training focused on multicultural issues than White clinicians. This last finding is quite tentative given that only 4 participants self-identified as Hispanic. Still, these results may suggest that acquiring training in multicultural issues evidences a selection bias such that certain clinical psychology trainees, namely

female and Hispanic trainees, are more interested in such issues and/or seek out additional training relevant to multicultural issues.

Limitations

The study findings need to be interpreted in light of several limitations. A primary concern regards the participant recruitment process. The obtained sample size was much lower than expected. As noted above, an *a priori* power analysis indicated a target sample size of 240 yet only 145 participants were recruited and took part in the study. Cell sizes across the 8 conditions ranged from as few as 13 participants up to 23 participants. This less-than-optimal sample size was the result of low response rates. The survey link was posted on multiple occasions to an email listserv for general clinical psychology as well as a listserv devoted specifically to forensic psychology. Additionally, members of the main forensic psychology professional organization (i.e., AP-LS) were emailed a solicitation to participate directly through that organization. Moreover, all solicitations included a request that recipients forward the survey link to colleagues. Despite these efforts the final sample fell far short of the target. Still, given medium to large effect sizes for the test presence/norm quality independent variable some findings were statistically significant. The recruitment issues limit the generalizability of these findings to the larger population of forensic evaluators. Moreover, it is unclear whether clinicians who belong to professional organizations or email listservs are representative of practicing clinicians. Additionally, participants in the sample almost exclusively identified as White, despite specific efforts in the solicitation email encouraging participation from racial and

ethnic minority clinicians. Thus, this demographic composition may reflect underrepresentation of these groups among doctoral-level forensic evaluators.

A concern is that given the smaller than expected sample and resultant decrease in statistical power small effects would not reach statistical significance. However, in many cases the race conditions differed very slightly, especially on ratings of norm quality where there was striking similarity across groups. Effect sizes (*d*) comparing race conditions ranged from 0.05 to 0.1. On a few comparisons there were small, though not statistically significant, effects. However, in these cases all but one indicated judgments contrary to the hypotheses. For example, on ratings of diagnostic confidence participants in the White youth condition expressed more confidence when presented with poor norms compared with participants in the African American youth condition ($d = 0.33$) and appeared to approach their judgment of African American youth slightly more conservatively. It was expected that given poor normative data clinicians would exhibit the opposite effect, that is, express more confidence in their decision when offering a diagnostic judgment for an African American youth when presented with lower quality test norms.

Further, no manipulation check was included so it is difficult to determine whether participants attended to the youth's race. Simply stating the youth's race in the vignette might not have been a strong enough manipulation to influence decision making or activate potential subtle racial biases. The conditions that included normative data attempted to cue participants to attend to the youth's race by stipulating that race was related to test scores on the hypothetical measure as well as to the diagnostic outcome. Inclusion of a questionnaire item asking participants to

identify the youth's race might have clarified whether or not participants noticed or attended to this piece of information. In contrast, the manipulation of the test presence/quality of norms elicited strong effects but this condition included multiple pieces of information that likely drew more attention from participants. Comparatively, the youth's race was mentioned once in the beginning of the case vignette and again, may not have been prominent enough to elicit an effect.

Implications and Future Directions

Overall this study is the first to examine the impact of youth race on psychologists' appraisal of norm quality. Thus, while these findings are encouraging they must be considered preliminary and exploratory. Still, the findings have implications for juvenile forensic mental health assessment. First, clinicians attended to the presence and quality of an assessment measure and adjusted their decision making practices accordingly. Given that clinicians reported that the quality of normative data decreased their diagnostic judgments as well as their confidence, perhaps a description or discussion of the normative data used in the context of applied assessments should be included in forensic reports. That is, forensic evaluators acknowledged that this feature of assessment measures influences their decision making. Thus, perhaps legal decision makers should be made aware of the characteristics and quality of test norms and any potential limitations that result.

Clinicians did not exhibit shifting standards regarding the quality of test norms for racial minority youth compared to White youth. Thus, there was no indication of a racial bias such that clinicians approached diagnostic judgments with more confidence and rated poor test norms as having higher quality when the evaluee was African

American compared to White. This is a positive finding for the field; however, this is the first study to examine this potential subtle racial bias and given the methodological limitations noted above, requires replication.

Future research in this area should seek to improve the recruitment strategy and acquire a larger, more racially and ethnically diverse sample of psychologists. This might be facilitated by the use of compensation or by acquiring postal mail addresses and attempting to collect hard-copy rather than electronic surveys. Additionally, future research should seek to examine this topic outside the context of juvenile forensic assessment and might consider a similar study among, for example, clinical neuropsychologists or in the context of more general child and adolescent mental health assessment.

It would be interesting to examine the main finding that presenting a psychological test influenced ratings on the dependent measures compared to the condition that did not include a test. Specifically, future research might examine if the presence of a psychological test per se influenced ratings or if it was simply the inclusion of additional information (i.e., the effect was not specifically due to the presence of a psychological test). Also, the salience of youth race might be increased by including additional stimulus material, perhaps including photographs of youth that differ only in skin tone. Further study might also expand the independent variables to include conditions for additional sociodemographic characteristics of youth, such as other racial or ethnic groups and the inclusion of a male youth condition. Lastly, it would be interesting to examine the potential for subtle racial or gender bias in juvenile forensic mental health assessment with the inclusion and manipulation of

additional youth characteristics such as prior arrest history and presence of callous-unemotional traits. Additional research in this area can be instructive in nature, aiming to address potential biases in juvenile forensic mental health assessment, to detect and attenuate them if present, and to provide recommendations to promote high-quality clinical practice.

APPENDICES

Appendix A

Case Vignettes

Participants in the No Test condition read this:

"Jessica is a 16 year old *(White/African American)* youth. She has been adjudicated delinquent in juvenile court. You have received a referral to assist the court in evaluating her for a potential substance abuse disorder. The frequency of substance abuse disorder in this setting is 50%. You complete a thorough file review and clinical interview but the diagnosis remains unclear."

All other conditions also read this (in addition to the information above):

"To try to clarify matters you administer a detailed, standardized test that assesses for substance abuse disorders in youth and is regularly used in this setting. The normative data for the assessment tool appears below. Age, gender, and race relate to test scores on this measure and the diagnostic outcome. This measure has a mean score of 100 and a standard deviation of 15, with the following cut-off scores: *120 to 129 - Borderline range* and *130 and above - Clinical range.* Jessica's standard score was 135."

Depending on the participant's condition the test's normative data was presented as follows:

	Ambiguous Norms	Small Sample Size	Poor Norms
Sample Size	Female Norm Sample – N = 120 Race/ethnicity: White youth – n=35 African American youth – n=35 Hispanic youth – n=35 Mixed or Other race – n=15	Female Norm Sample – N = 35 Race/ethnicity: White youth – n=10 African American youth – n=10 Hispanic youth – n=10 Mixed or Other race – n=5	Female Norm Sample – N = 35 Race/ethnicity: White youth – n=10 African American youth – n=10 Hispanic youth – n=10 Mixed or Other race – n=5
Sample Age	Age group 15 to 17	Age group 15 to 17	Age group 17 to 19
Obsolescence	Test most recently normed in 2005	Test most recently normed in 2005	Test most recently normed in 1985

Table 1

Summary of Sample Demographic Characteristics

Sex	
Female	40.7%
Male	59.3%
Race/Ethnicity	
American Indian/Alaskan Native	1.5%
Asian	1.5%
Hispanic/Latino	3.0%
White	91.7%
Multiracial	2.3%
Highest Degree	
Ph.D.	62.7%
PsyD	23.1%
Ed.D	3.0%
Master's Degree	11.2%
Years since graduation *M*(*SD*)	17.25 (12.04)
Years assessment experience *M*(*SD*)	20.21 (11.28)
Years forensic experience *M*(*SD*)	15.99 (10.70)
Median assessments in past year	49.5
Percent minority evaluees *M*(*SD*)	48.38 (27.93)
Median hours multicultural didactics	75
Median hours multicultural supervision	25
% Board certified forensic psychology	20.3%

Table 2

ANOVA Source Table

Dependent Variable	Source	*df*	*SS*	*MS*	*F*	*p*
Likelihood	Race	1	670.07	670.07	2.11	.15
	Test	3	13,355.13	4,451.71	14.0	<.001
	Race x Test	3	1,555.38	518.50	1.63	.19
	Error	137	43,557.19	317.94		
	Total	145	505,297			
Confidence	Race	1	0.59	0.59	0.27	.61
	Test	3	84.92	28.31	12.66	<.001
	Race x Test	3	3.05	1.02	0.45	.72
	Error	137	306.45	2.24		
	Total	145	1,634			
Norm Quality	Race	1	0.06	0.06	0.03	.86
	Test	2	65.79	32.90	16.49	<.001
	Race x Test	2	0.25	0.12	0.06	.94
	Error	98	195.46	1.99		
	Total	104	952			

Table 3

Descriptive Statistics for Full Factorial Design

Dependent Variable	No Test		Ambiguous Norms		Small Sample Size		All Bad Norms	
	White *n*=18	African American *n*=23	White *n*=13	African American *n*=16	White *n*=18	African American *n*=21	White *n*=17	African American *n*=19
Likelihood *M*(*SD*)	43.83(13.27)	39.83(19.91)	69.08(9.38)	68.06(16.06)	59.06(22.59)	61.62(15.27)	63.18(21.15)	48.21(18.88)
Confidence *M*(*SD*)	1.94(1.06)	2.22(1.45)	4.15(1.63)	3.88(1.67)	3.50(1.54)	3.48(1.60)	2.65(1.69)	2.16(1.30)
Norm Quality *M*(*SD*)	-	-	3.85(1.41)	3.69(1.85)	2.50(1.38)	2.43(1.60)	1.71(0.99)	1.79(1.08)

Table 4

Summary of Post-hoc Analyses for the Main Effect of Test Presence/Norm Quality

Dependent Variable	Test Presence/Norm Quality Condition			
	No Test	All Bad Norms	Small n	Ambiguous Norms
Likelihood of Disorder *M(SD)*	41.56 (17.24)[abc]	55.28 (21.10)[ad]	60.44 (18.78)[b]	68.52 (13.27)[cd]
Confidence *M(SD)*	2.10 (1.28)[abc]	2.39 (1.50)[ad]	3.49 (1.55)[b]	4.0 (1.63)[cd]
Norm Quality *M(SD)*	-	1.75 (1.03)[ab]	2.46 (1.48)[a]	3.76 (1.64)[b]

Note. Means with the same superscript were significantly different at $p = .05$

Table 5

Nonparametric Correlations (Spearman's Rho) Between Training and Experience Variables

	1	2	3	4	5	6
1. Years since degree	-					
2. Years assessment experience	.936**	-				
3. Years forensic assessment experience	.853**	.892**	-			
4. Percentage assessments involving minority clients	-.202*	-.184*	-.170	-		
5. Hours didactic training focused on multicultural issues	-.333**	-.329**	-.280**	.266**	-	
6. Hours supervision focused on multicultural issues	-.217*	-.235**	-.240**	.364**	.678**	-
7. Self-appraised multicultural competence	-.021	.018	-.001	.277**	.282**	.231**

*p<.05 **p<.0

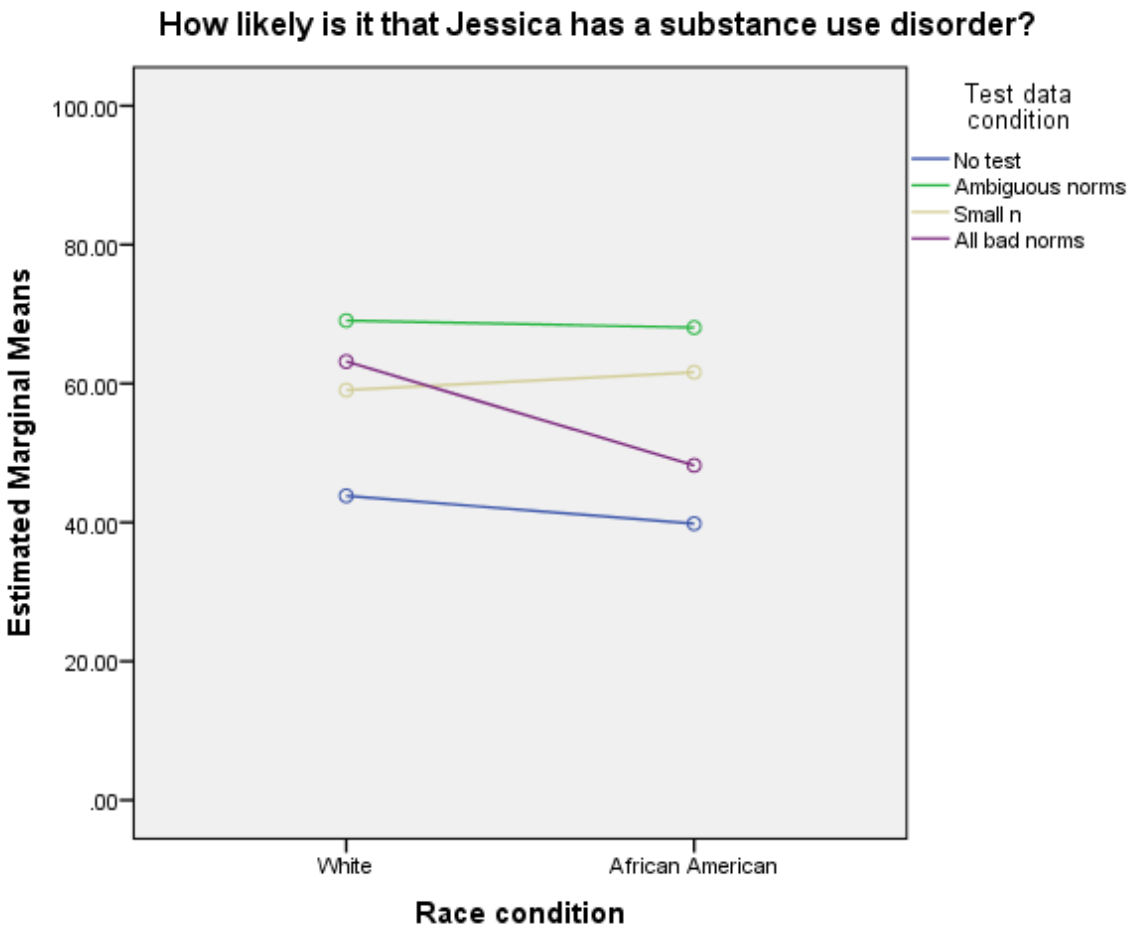

Figure 1. Means plot for ratings of disorder likelihood.

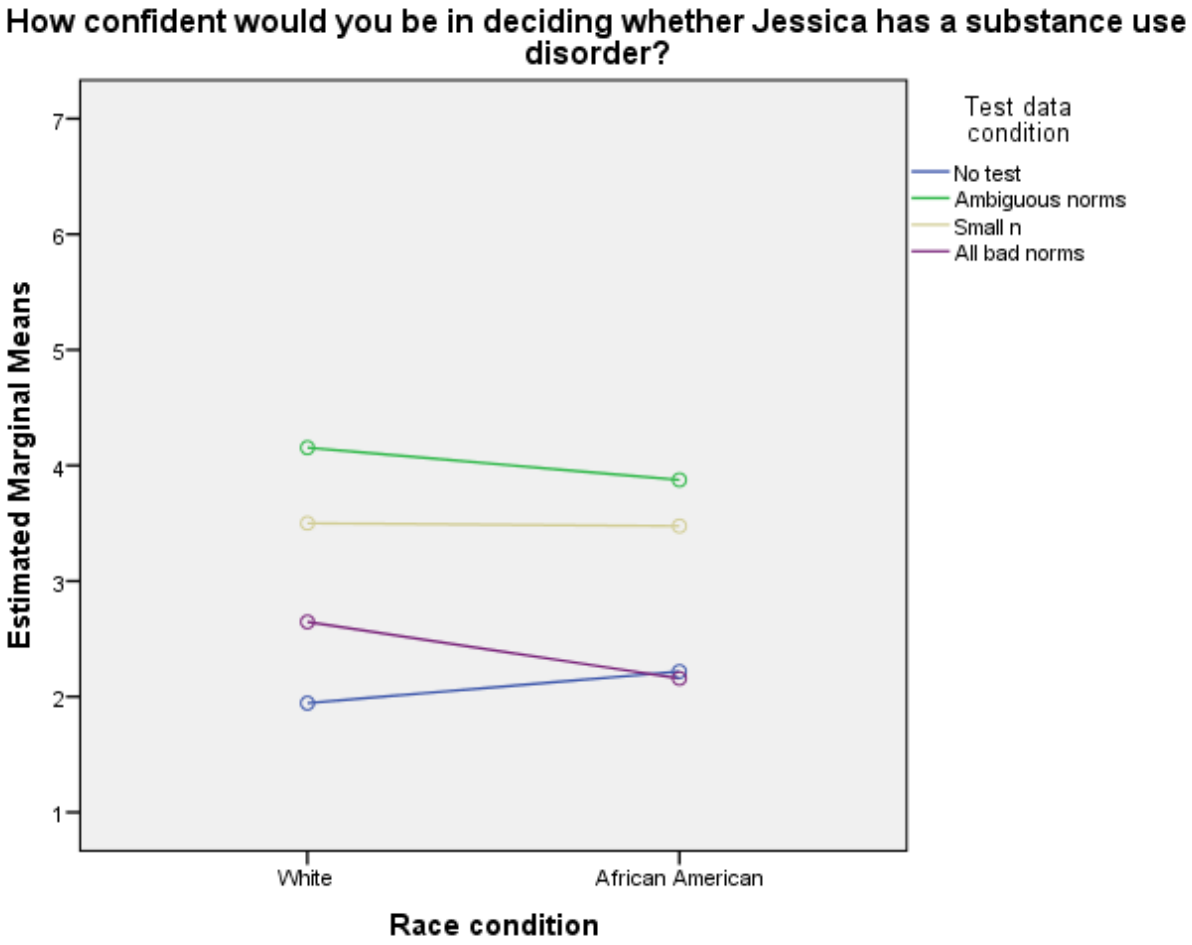

Figure 2. Means plot for ratings of judgment confidence.

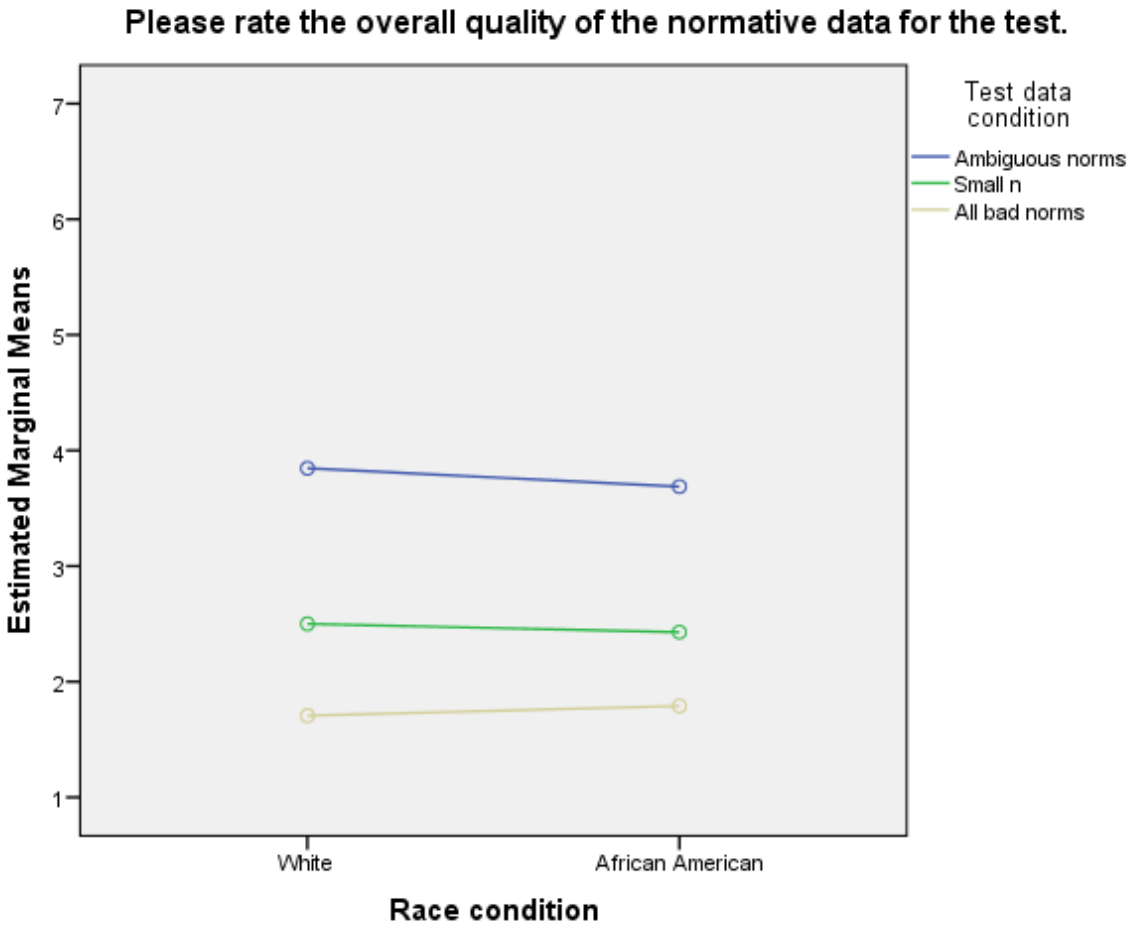

Figure 3. Means plot for ratings of norm quality.

BIBLIOGRAPHY

Abram, K. M., Teplin, L. A., McClelland, G. M., & Dulcan, M. K. (2003). Comorbid psychiatric disorders in youth in juvenile detention. *Archives of General Psychiatry, 60,* 1097-1108. doi: 10.1001/archpsyc.60.11.1097

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychological Association. (2002). American Psychological Association ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060-1073. doi: 10.1037/0003-066X.57.12.1060

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87,* 84-94. doi: 10.1207/s15327752jpa8701_07

Biernat. M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology, 66,* 5-20. doi: 0022-3514/94

Breda, C. S. (2003). Offender ethnicity and mental health service referrals from juvenile courts. *Criminal Justice and Behavior, 30,* 644-667. doi:10.1177/0093854803256451

Bridges, A. J., & Holler, K. A. (2007). How many is enough? Determining optimal sample sizes for normative studies in pediatric neuropsychology. *Child Neuropsychology, 13,* 528-538. doi: 10.1080/09297040701233875

Bright, C. L., & Jonson-Reid, M. (2008). Onset of juvenile court involvement: Exploring gender-specific associations with maltreatment and poverty. *Children and Youth Services Review, 30,* 914-927. doi:10.1016/j.childyouth.2007.11.015

Carter, R. T., & Forsyth, J. M. (2007). Examining race and culture in psychology journals: The case of forensic psychology. *Professional Psycholgy: Research and Practice, 38,* 133-142. doi: 10.1037/0735-7028.38.2.133

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New Jersey: Lawrence Erlbaum.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science,243,* 1668–1774. doi: 10.1126/science.2648573

Echemendia, R. J., & Harris, J. G. (2004). Neuropsychological test use with Hispanic/Latino populations in the United States: Part II of a national survey. *Applied Neuropsychology, 11,* 4-12. doi: 10.1207/s15324826an1101_2

Faust, D. (1984). *Limits of scientific reasoning.* Minneapolis, MN: University of Minnesota Press.

Faust, D. (2005). Why Paul Meehl will revolutionize the philosophy of science and why it should matter to psychologists. *Journal of Clinical Psychology, 61,* 1355-1366. doi:10.1002/jclp.20185

Faust, D. (2008). Why meta-science should be irresistible to decision researchers. In I. J. Krueger (Ed.), *Rationality and social responsibility: Essays in honor of Robyn Mason Dawes* (pp.91-110). New York, NY: Psychology Press.

Faust, D., Ahern, D. C., & Bridges, A. J. (2012). Neuropsychological (brain damage) assessment. In, D. Faust, *Coping with psychiatric and psychological testimony* (pp.363-469). New York, NY: Oxford University Press.

Faust, D., Ziskin, J., & Hiers, J. B. (1991). Brain damage claims: Coping with neuropsychological evidence. Venice, CA: Law & Psychology Press.

Garb, H. N. (1997). Race bias, social class bias, and gender bias in clinical judgment. *Clinical Psychology: Science and Practice, 4*, 99-120.

Geib, C. F., Chapman, J. F., D'Amaddio, A. H., & Grigorenko, E. L. (2011). The education of juveniles in detention: Policy considerations and infrastructure development. *Learning and Individual Differences, 21*, 3-11. doi:10.1016/j.lindif.2010.05.002

Grisso, T. (1996). Society's retributive response to juvenile violence: A developmental perspective. *Law and Human Behavior*, *20*,229-247. doi:10.1007/BF01499022

Grisso, T. (1998). *Forensic evaluation of juveniles.* Sarasota, FL, US: Professional Resource Press.

Grisso, T. (2005a). Evaluating the properties of instruments for screening and assessment. In, Grisso, T., Vincent, G., & Seagrave, D. (Eds.). *Mental Health Screening and Assessment in Juvenile Justice* (pp. 71-97). New York: Guilford Press.

Grisso, T. (2005b). Why we need mental health screening and assessment in juvenile justice programs. In, Grisso, T., Vincent, G., & Seagrave, D. (Eds.). *Mental Health Screening and Assessment in Juvenile Justice* (pp. 3-21). New York: Guilford Press.

Gushue, G. V. (2004). Race, color-blind attitudes, and judgments about mental health: A shifting standards perspective. *Journal of Counseling Psychology, 51,* 398-407. doi:10.1037/0022-0167.51.4.398

Hawkins, D. F., Laub, J. H., Lauritsen, J. L., Cothern, L. (2000). Race, ethnicity, and serious and violent juvenile offending. Washington, DC: US Department of Justice.

Hoge, R. D. (2008). *Assessment in juvenile justice systems*. In Hoge, R. D., Guerra, N. G., & Boxer, P. (Eds.) *Treating the juvenile offender* (54-75). New York, NY, US: Guilford Press.

Kalechstein, A. D., van Gorp, Wilfred G., & Rapport, L. J. (1998). Variability in clinical classification of raw test scores across normative data sets. *Clinical Neuropsychologist, 12,* 339-347. doi:10.1076/clin.12.3.339.1991

Kataoka, S. H., Zima, B. T., Dupre, D. A., Moreno, K. A., Yang, X., & McCracken, J. T. (2001). Mental health problems and service use among female juvenile offenders: Their relationship to criminal history. *Journal of the American Academy of Child & Adolescent Psychiatry, 40,* 549-549-555. doi:10.1097/00004583-200105000-00014

Kempf-Leonard, K. (2007). Minority youths and juvenile justice: Disproportionate minority contact after nearly 20 years of reform efforts. *Youth Violence and Juvenile Justice, 5*, 71-87. doi:10.1177/1541204006295159

Koocher, G. P. (2006). Ethical issues in forensic assessment of children and adolescents. In Sparta, S. N., & Koocher, G. P. (Eds.). *Forensic mental health assessment of children and adolescents* (pp.46-63). New York: Oxford University Press.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194-216. doi: 10.1037/h0048070

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). Oxford: Oxford University Press.

National Council on Crime and Delinquency. (2007). *And justice for some: Differential treatment of youth of color in the justice system.* Retrieved, November 9, 2010 from http://www.nccd-crc.org/nccd/pubs/2007jan_justice_for_some.pdf

Otnow Lewis, D., Balls, D. A., & Shanok, S. S. (1979). Some evidence of race bias in the diagnosis and treatment of the juvenile offender. *American Journal of Orthopsychiatry, 49*, 53-61. doi: 10.1111/j.1939-0025.1979.tb02585.x

Otto, R. K., Borum, R., & Epstein, M. (2012). Evaluation of children in the juvenile justice system. In, D. Faust, *Coping with psychiatric and psychological testimony* (pp.754-772). New York, NY: Oxford University Press.

Patel, V., Flisher, A. J., Hetrick, S., & McGorry, P. (2007). Mental health of young people: A global public-health challenge. *Lancet, 369,* 1302-1313. doi: 10.1016/S0140-6736(07)60368-7

Piquero, A. R. (2008). Disproportionate minority contact. *The Future of Children, 18*(2), 59-79. doi:10.1353/foc.0.0013

Powell, M. B., & Bartholomew, T. (2003). The treatment of multicultural issues in contemporary forensic psychology textbooks. *Psychiatry, Psychology, and Law, 10,* 254-261.

Rawal, P., Romansky, J., Jenuwine, M., & Lyons, J. S. (2004). Racial differences in the mental health needs and service utilization of youth in the juvenile justice system. *The Journal of Behavioral Health Services & Research, 31*(3), 242-254. doi: 10.1097/00075484-200407000-00002

Shufelt, J., & Cocozza, J. (2006). *Youth with mental health disorders in the juvenile justice system: Results from a multi-state prevalence study.* Delmar, NY: National Center for Mental Health and Juvenile Justice. Retrieved from http://www.ncmhjj. com/pdfs/publications/PrevalenceRPB.pdf

Smedley, B., Stith, A., & Nelson, A. (2003). *Unequal treatment: Confronting racial and ethnic disparities in health care.* Washington, DC: National Academy Press.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests.* New York, NY: Oxford University Press.

Teplin, L. A., Abram, K. M., McClelland, G. M., Dulcan, M. K., & Mericle, A. A. (2002). Psychiatric disorders in youth in juvenile detention. *Archives of General Psychiatry, 59*, 1133-1143. doi: 10.1001/archpsyc.59.12.1133

Teplin, L. A., Abram, K. M., McClelland, G. M., Washburn, J. J., & Pikus, A. K. (2005). Detecting mental disorder in juvenile detainees: Who receives services. *American Journal of Public Health, 95*, 1773-1780.

U.S. Census Bureau. (2008). *2008 national population projections.* Retrieved, November 10, 2010, from http://www.census.gov/population/www/projections/2008projections.html

Viljoen, J. L., McLachlan, K., & Vincent, G. M. (2010). Assessing violence risk and psychopathy in juvenile and adult offenders: A survey of clinical practice. *Assessment, 17*, 377-395. doi:10.1177/1073191109359587

Vincent, G. M., Chapman, J., & Cook, N. E. (2011). Risk-needs assessment in juvenile justice: Predictive validity of the SAVRY, racial differences, and the contribution of needs factors. *Criminal Justice and Behavior, 38*(1), 42-42-62. doi: 10.1177/0093854810386000

Vincent, G. M., Grisso, T., Terry, A., & Banks, S. (2008). Sex and race differences in mental health symptoms in juvenile justice: The MAYSI-2 national meta-analysis. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*, 282-290. doi: 10.1097/CHI.0b013e318160d516

Whaley, A. L., & Geller, P. A. (2007). Toward a cognitive process model of ethnic/racial biases in clinical judgment. *Review of General Psychology, 11*, 75-96. doi:10.1037/1089-2680.11.1.75

Wood, J. M., Garb, H. N., & Nezworski, M. T. (2007). Psychometrics: Better measurement makes better clinicians. In S. O. Lilienfeld, & W. T. O'Donohue (Eds.), *The great ideas of clinical science: 17 principles that every mental health professional should understand* (pp.77-92). New York, NY: Routledge.