The Question of Cultural Bias in Assessment and Diagnosis of Ethnic Minority Clients: Let's Reject the Null Hypothesis

Robert G. Malgady
New York University and Fordham University

Cultural bias in the mental health assessment and psychiatric diagnosis of ethnic minorities has been asserted and challenged on the basis of clinical impression and a variety of psychometric criteria. The lack of a definitive empirical basis to resolve issues of assessment and diagnostic bias means that the null hypothesis (no bias or cross-cultural uniformity) prevails. This article argues that the traditional hypothesis to be nullified should be challenged. The consequences associated with its incorrect retention (Type II error = disservice to minority clients) may be more serious than its incorrect rejection (Type I error = misdirection to the mental health service system). If a client-centered error is judged more serious than a service system error, then a statement of bias or cross-cultural variance should be the null hypothesis until such time as empirical data suggest otherwise.

The presumptions of valid measurement of psychological symptomatology and accurate diagnosis of psychiatric disorders are fundamental to the integrity of mental health research and to the delivery of effective mental health services. The fourth edition of the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM-IV; 1994) reflects an increasing recognition of the importance of considering the cultural diversity of clients in rendering psychiatric diagnoses. Yet limited empirical data and even sparser theoretical preconceptions hinder researchers, professional practitioners, and policymakers in deciding whether or not—and, if so, how—a culturally informed mental health assessment ought to take place (Lopez, 1988; Lopez & Hernandez, 1987; Malgady, 1994). This issue is of special concern because effective psychopharmacological and psychotherapeutic treatment planning and continuity of care are predicated, in part, on accurate evaluation of symptom severity and diagnosis.

The issue of uniformly valid assessment has far reaching implications in many domains. In psychiatric epidemiological research, estimates of accurate rates of mental health disorders depend on the administration of symptom scales, such as the widely used National Institute of Mental Health (NIMH) Center for Epidemiological Studies—Depression Scale (CES-D; Radloff, 1977), which are uniformly valid in different ethnic populations. Similarly, nationwide NIMH Epidemiological Catchment Area studies (e.g., Kessler et al., 1994) have been based on the administration of the Diagnostic Interview Schedule (DIS), a structured psychiatric interview targeting DSM-III-R diagnostic criteria (Robins, Helzer, Costello, & Regier, 1981). However, the comparison of prevalence rates across ethnic groups is appropriate only if the diagnostic criteria and the algorithms that bind such criteria to the psychiatric nosology are uniformly applicable across these groups (Malgady & Rogler, 1993; Shroff et al., 1995). The cultural considerations introduced for the first time into the DSM-IV (e.g., "ethnic and cultural considerations," p. xxiv; glossary of culture bound syndromes), epidemiological research indicating the need to adjust both the inclusion criteria for symptoms and the DIS algorithms cross-culturally (Canino et al., 1987), and other research indicating higher prevalence and comorbidity rates of selected DSM-III-R disorders in certain ethnic minority populations (e.g., Kessler et al., 1994), together cast suspicion on the presumption of cross-cultural uniformity of assessment.

Uniformity is especially problematic given the more rapid increases in ethnically diverse populations in this country relative to the growth rate of the non-Hispanic White population. According to one source, it is estimated that about 40% of the clients in the mental health service system will be members of ethnic minority groups in the year 2000 (Cross, Bazron, Dennis, & Isaacs, 1989). Thus, it is conceivable that the professional psychologist's array of assessment techniques may indeed need to vary in accordance with the specific ethnic group identity, cultural predispositions, and language of his or her clientele.

Evidence of Assessment and Diagnostic Bias

In the tradition of Sir Ronald Fisher, the prevailing view representing the status quo leads to the formulation of a null hypothesis of uniformity of assessment (i.e., no difference between ethnic groups). This hypothesis prevails unless sufficient empirical force enables its nullification or rejection. The competing hypothesis of cultural variance in assessment (i.e.,

Robert G. Malgady received his PhD from the University of Tennessee in 1975. He is currently professor of quantitative studies at New York University and is senior research associate at the Hispanic Research Center, Fordham University. He teaches and writes on cultural sensitivity in measurement and psychotherapy and on psychometric and statistical methods.

This research was supported in part by National Institute of Mental Health Grant R01MH30569 from the Services Research Branch of the Division of Applied and Services Research.

I wish to thank Lloyd H. Rogler and Steven R. Lopez for their helpful comments on this article.

Correspondence concerning this article should be addressed to Robert G. Malgady, 10 White Lion Drive, Montrose, New York 10548.
Worobey, 1989) and as a function of single- and coupled-
factor invariance between White, Black, Mexican American, and Puerto Rican groups. Unfortunately, some studies report factor invariance between various psychological tests and on diagnostic rates of disorder. Bilingual patients speaking their second language evidence nullifies the status quo. If the null hypothesis prevails, the assessment or diagnostic technique in question is said to be unbiased, whereas if the null hypothesis is rejected, the technique is considered culturally biased.

Unfortunately, in the ethnic minority mental health literature, the empirical data needed to render a decision on this issue are either lacking or equivocal (e.g., Costantino & Malgady, in press; Lopez, 1988; Sue, 1988). Thus, with few exceptions, the presumption of uniformity across groups has prevailed.

At the most rudimentary level, polemics persist about apparent bias in the symptom indicators and diagnostic criteria defining psychopathology in tests and interview schedules developed from the perspective of mainstream American society. The typical argument is that some items in widely used standardized tests, such as the Minnesota Multiphasic Personality Inventory (MMPI), refer to culturally patterned behaviors, beliefs, and feelings that are not pathological in certain cultures (Padilla & Ruiz, 1975; Rogler & Hollingshead, 1985). Others in cross-cultural psychiatry have warned about the danger of ethnocentrism, particularly regarding the category fallacy, a term that refers to the imposition of Western nosologies of psychopathology, such as the DSM-III, on members of other cultures (Kleinman & Good, 1985).

Probably the most attention concerning psychometric definitions of test bias has been paid to mean differences on standardized tests and the attendant question of whether ethnically specific norms need to be developed (e.g., Gynther & Green, 1980; Pritchard & Rosenblatt, 1980). The extensive empirical research on normative differences between ethnic populations on various psychological tests and on diagnostic rates of disorder is mostly equivocal (Lopez, 1988).

Measurement bias also has been defined in terms of latent factor structure of tests across different populations. Factor invariance has been examined among Whites, Blacks, and Hispanics in the intelligence testing literature (e.g., Gutkin & Reynolds, 1981a, 1981b), but little is known about cross-cultural variations in personality tests or symptom scales. One exception is the CES-D, which has been factor analyzed among White, Black, Mexican American, and Puerto Rican groups. Unfortunately, some studies report factor invariance between ethnic groups (Aneshensel, Clark, & Frerichs, 1983; Roberts, 1980), whereas other studies have reported distinct factor structures among ethnic minority women (Guarnaccia, Angel, & Worobey, 1989) and as a function of single- and coupled-headed household type (Stroup-Benham, Lawrence, & Trevino, 1992). Similarly, research on the MMPI has produced ambiguous findings. Differences in the number and composition of MMPI factors among Whites, Blacks, and Mexican Americans have been reported in some studies (Holland, 1979), whereas other studies have reached the opposite conclusion (Prewitt-Diaz, Nogueras, & Draguns, 1984).

Other definitions of bias refer to population differences in how test scores relate to an external criterion-related measure. Differential validity is a question of equivalence across populations in terms of correlation coefficients; differential prediction is a question of equivalence of the accompanying regression equations; and measurement equivalence refers to how observed measurements map onto latent traits of members of different populations (Drasgow, 1984). Evidence that the validity of a particular assessment technique varies across ethnic groups would constitute evidence of test bias, which would imply that assessments are less clinically relevant for some ethnic group. Evidence of differential regression equations would suggest that bias takes the form of under- or overprediction of outcome, which would suggest systematic unfairness in clinical disposition. When latent measurement models are not equivalent across ethnic groups, however, bias occurs because members of the different groups with the same underlying trait value (e.g., actual symptom severity) receive different observed assessments.

Drasgow (1984) showed that without measurement equivalence significance tests are nearly powerless to detect true differences between validity coefficients; thus, the failure to observe significant differences in validity is not evidence that tests are fair to ethnic minority groups. Ellis (1989) and her colleagues (Ellis, Becker, & Kimmel, 1993; Ellis & Kimmel, 1992) applied this technique to compare English, German, and French translations of items on a personality inventory and a mental health attitude scale. They found several items to be in need of revision or deletion. Ellis (1989) concluded that “even the most meticulous and painstaking translation and back-translation will not ensure measurement equivalence” (p. 919).

Thus, once again we reach an inescapable conclusion: There is only sporadic psychometric evidence of bias or lack of it in cross-cultural studies of issues pertaining to validity in psychological assessment. This evidence varies depending on the ethnic and racial groups being compared, the assessment techniques in question, and by the very psychometric definition of test bias itself (Costantino & Malgady, in press).

In unstructured situations, such as routine psychiatric interviews, an ethnic minority client is implicitly compared to the clinician’s generalized perception of psychopathology relative to DSM-IV diagnostic criteria. Problems arise in diagnostic situations when researchers inquire about how cultural variables might influence the validity of qualitative clinical judgments. For instance, research with bilingual people suggests that greater psychopathology is inferred from psychiatric interviews conducted in the patient’s first language than in the second language, yet other studies have reached the opposite conclusion (Vazquez, 1982).

Marcos (1994) demonstrated how process variables in the client–diagnostician interaction affect the diagnostic impression. Bilingual patients speaking their second language evidence a variety of speech and motor disturbances. Marcos called attention to reduced self-disclosure, flattened affect, hesitations, stilted verbal expressions, and, among other behaviors, kinesic behavior indicative of poor sense of self and emotional disturbances.

However, the relatively few research studies of severity of psychopathology involving bilingual participants have been based on very small samples, have been fraught with methodological confounds, and have led to equivocal conclusions. The equally scarce literature on diagnostic outcomes is based largely on clinical impression.

Despite several decades of rhetoric and scattered research efforts, there is little empirical or theoretical basis for consensus on whether or not (or how) cultural variables affect the out-
comes of standardized assessments and psychiatric diagnosis. The research does not provide sufficient evidence—because of inconsistent findings, flawed findings, or no findings—to contradict the status quo presumption that psychological measurement techniques are uniformly, cross-culturally valid.

The Null Hypothesis in Cross-Cultural Assessment and Diagnosis

The usual null hypothesis in statistical inference poses a "no difference" proposition that prevails as the status quo governing our actions until such time as empirical evidence enables its rejection. Thus, traditional statistical reasoning leads quite naturally to posing uniformity of assessment (or, no bias) as the null hypothesis in cross-cultural research, regardless of the particular psychometric definition of bias.

However, there is no logical reason why other nonzero versions of a null hypothesis cannot be formulated. As Glass and Hopkins (1970) stated, the null hypothesis "is simply a numerical statement about an unknown parameter . . . or a difference between parameters" (pp. 203–204). Hays (1973) stated that the null hypothesis "can specify any of the possible values for one or more parameters, and this use of the word ‘null’ is only incidental [to the zero value usually specified]" (p. 369). Moreover, a recent article by Cohen (1994) severely criticized the traditional formulation of a null hypothesis as zero difference, which he calls the "nil" hypothesis.

An equally tenable null hypothesis in cross-cultural research on assessment and diagnostic bias could be that there is a difference (or degree of bias) between ethnic groups that is judged to be clinically meaningful in magnitude. In other words, the magnitude of the difference would be the lowest value or threshold that would result in a change in the clinical disposition of individual clients in the mental health service system. Such nonzero null hypotheses might be posed either by clinical judgment or in terms of Cohen's (1977) conventions in behavioral science for small, moderate, and large "effect sizes." The competing or alternate hypothesis, then, would not be that the cultural difference is zero, but rather that it is of a magnitude less than the specified threshold of clinical meaningfulness. In other words, the alternate hypothesis is simply that the difference is of an inconsequential magnitude.

Given the current status of empirical research on assessment of ethnic minorities, the null hypothesis of no cross-cultural bias prevails in the presence of equivocal data, poor data, or no data. Statisticians tell us that the null hypothesis can never be proved; nonetheless, as long as it prevails, it governs our actions. Conversely, had the nontraditional null hypothesis of cultural bias (i.e., a clinically meaningful cross-cultural difference) been posed, then this would have prevailed given the current circumstances. In order to reconsider which null hypothesis ought to represent the status quo, the consequences of errors attending the decision to adopt either hypothesis must be evaluated. This decision, I argue, is fundamentally a value judgement.

The Consequences of Type I and Type II Errors

According to conventional standards of acceptable margins of risk, Type I errors (incorrectly rejecting the null hypothesis) usually are permitted a maximum probability of .05, whereas Type II errors (failure to reject a false null hypothesis) are afforded a maximum probability of .20. In the psychological literature today, the former convention is rather strictly enforced, but the latter is not. It is rare to find examples of research in the literature in which the probability of Type II error (or, conversely, statistical power) is reported. Cohen (1994) estimated that the prevalence of Type II errors is about .50 in practice. Ideally, the probabilities of both types of errors should be as small as possible, of course, but according to the conventional minimum standards of risk and common practice, Type I errors are valued as more grievous than Type II errors (Hays, 1973).

Which type of error is effectively treated as more grievous, therefore, is dependent entirely on which of two competing hypotheses are formulated as the null hypothesis. Consequently, if, first, a determination is made regarding which of the two types of statistical decision errors is more serious (i.e., false conclusion of cultural uniformity or no bias vs. false conclusion of cultural variance or bias), then the corresponding null hypothesis can be formulated such that the more grievous type of error is fixed as Type I. Consequently, the more serious error will be minimized. Hays (1973) actually advocated this procedure. Taking the most grievous error of the two types, the corresponding hypothesis that risks this error "or its statistical equivalent is cast in the role of the null hypothesis . . . so that the abhorrent Type I error is very unlikely to be committed" (p. 368).

In the usual case where the null hypothesis specifies no difference between groups (see Table 1), a Type I error would mean that assessment and diagnostic procedures are judged to be culturally biased when, in fact, they are not; this scenario would mean that new culture-specific modifications would need to be developed but would be used in vain. Thus, if a Type I error is committed and these new services do not really improve the mental health care of ethnic minorities, significant resources will have been misspent. In this case, the burden of error will be shouldered by the mental health system. However, if no bias is incorrectly concluded (a Type II error—see Table 1), minorities would end up being unfairly evaluated by standardized assessment techniques, would be misdiagnosed, and

<table>
<thead>
<tr>
<th>Consequence</th>
<th>Type I error</th>
<th>Type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td>Uniformity</td>
<td>Difference</td>
</tr>
<tr>
<td>Decision</td>
<td>Difference</td>
<td>Uniformity</td>
</tr>
<tr>
<td>Burden</td>
<td>Mental health system</td>
<td>Minority client</td>
</tr>
<tr>
<td>Value judgment</td>
<td>Less serious</td>
<td>More serious</td>
</tr>
</tbody>
</table>

Table 1
Comparison of Consequences of Type I and Type II Errors for Traditional versus Nontraditional Null Hypotheses

Traditional null hypothesis: Cross-cultural uniformity

Nontraditional null hypothesis: Cultural variance

<table>
<thead>
<tr>
<th>Consequence</th>
<th>Type I error</th>
<th>Type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td>Difference</td>
<td>Uniformity</td>
</tr>
<tr>
<td>Decision</td>
<td>Uniformity</td>
<td>Difference</td>
</tr>
<tr>
<td>Burden</td>
<td>Minority client</td>
<td>Mental health system</td>
</tr>
<tr>
<td>Value judgment</td>
<td>More serious</td>
<td>Less serious</td>
</tr>
</tbody>
</table>
ultimately would possibly receive less than effective treatment plans.

To the contrary, should cultural variance or bias be specified as the null hypothesis, the two types of errors would be reversed (see Table 1). Client-centered errors would become the more serious Type I errors, and system-centered errors would become Type II errors. In other words, mainstream assessment procedures would be presumed guilty of bias against other cultures unless proven innocent.

Value Judgment in Choice of Type I Error

Which type of error is more egregious rests ultimately on a value judgment. Hays (1973) provided a lucid analogy in his introduction to hypothesis testing using principles of decision theory. After calculating the probability of the two types of errors, he stated that the choice of which error to risk “depends upon the value... of the various outcomes” (p. 344). Using a Bayesian decision-outcome model, he illustrated how to quantify the value of the two errors in terms of a concept known in economic decision making as expected loss, and he demonstrated methods for choosing decision rules that minimize expected loss (Hays, 1988). However, such theoretic decision methods are premised on the assignment of a loss value to the errors, which is simpler when the outcome is financial. Hays pointed out that these methods fail in psychological research precisely because the two types of errors are often qualitatively different. He provided an example that reiterates the present dilemma of choosing between person and society: cost to the community of administering a diagnostic method for potential suicides versus occurrence of suicide.

Because the mental health service system is, by definition, intended to serve mentally ill individuals and because minorities comprise a significant and growing clientele within this service system, errors of disservice to clients would seem to be more grievous than errors of misdirection to the service system. If this value judgment is compelling, Type II errors would be more serious than Type I errors for the minority client when the zero-difference null hypothesis is the prevailing status quo. Therefore, this null hypothesis should presently be rejected a priori.

There are two arenas in which the traditional null hypothesis should be rejected: research and clinical practice. Researchers investigating ethnic bias in psychological tests or psychiatric diagnosis should formulate a nonzero null premise for statistical hypothesis testing. By the same token, this same presumption should be implemented in clinical practice. When assessing an ethnic minority client, clinical practitioners should not interpret the results of standardized psychological tests as if they are equivalent to the White middle-class norm unless such instruments have been scrutinized for a variety of types of bias within the client’s particular ethnic group. Because so few psychological tests have demonstrated measurement equivalence, the most fundamental definition of bias, clinicians should be prejudicial in interpreting psychological test results of ethnic minority clients. This presumption that psychological tests have an element of ethnic bias increases the clinician’s uncertainty regarding the “true” assessment of a client’s psychological functioning. This uncertainty will be attenuated in situations in which multiple assessment procedures that are based on diverse methodologies (such as objective inventories, projective techniques, and face-to-face interviews) provide a convergent clinical profile. When the multiple assessment procedures do not converge, however, the client’s assessment should remain highly suspect. Ultimately, the disposition of the client rests on clinical judgment, but it is hoped that this judgment is culturally informed and not adumbrated by standardized test data. Clinicians should also be alert to errors of clinical disposition that result from biased assessment early in the treatment process.

Increased attention to human diversity is evident in the American Psychological Association’s revised Standards for Educational and Psychological Testing (1985), which includes numerous cautions against using norms, item content, and validity and prediction estimates with demographically, linguistically, and culturally different populations. Similarly, the DSM-IV imposes cultural boundaries on many psychiatric disorders that were previously conceived of as biologically invariant. The spirit of these caveats is consistent with the notion of abandoning the traditional null hypothesis and proceeding as if testing and diagnostic procedures are not routinely valid with members of diverse ethnic populations.

This course of action is not without drawbacks. The linguistic, cultural, and demographic diversity of many ethnic minority groups has been widely documented (Rogler, Malgady, & Rodriguez, 1989; Malgady, 1994). The diversity of subcultures within ethnic groups poses complex problems for proponents of ethnic-specific norms for psychological tests and culture-specific diagnostic procedures or criteria. If the cultural variance null hypothesis were to become the standard proposition in the absence of compelling empirical answers, separate test forms (differing in item content, language idioms, and norms) and diagnostic criteria might be required for each distinct minority subculture. If researchers affirmed that standardized tests and diagnostic criteria require separate specification, because ethnic group differences are significant and not trivial, the consequence would be a menu of test forms and associated normative reference data and variations of special considerations of diagnostic criteria that depend on the client’s particular combination of ethnic group membership, language proficiencies, and demographic background (to say nothing of the usual structuring of norms by gender and by age distribution).

As philosophers of science have long observed, the rigor of the scientific method is no stranger to human value judgment. Science is not insulated from value judgments at the level of verification of theory. Four decades ago, Rudner (1954/1961) argued that value judgments are logically involved in the validation of scientific hypotheses and that value judgment is a “requisite step” (p. 31) in the process. The very decision, for instance, of choosing a particular significance level, such as the conventional .05 level, or a particular level of statistical power involves an irrevocable value judgment. Rudner further stressed that unless the scientist openly confronts the intrinsic nature of value judgments in the search for objectivity, values decisions will be made “intuitively, unconsciously, and haphazardly” (p. 34) along the way.

References


NULL HYPOTHESIS

77