In recent years, psychologists studying children's development have become increasingly concerned about issues of cultural sensitivity. Although this heightened awareness is encouraging, there is still considerable confusion and lack of clarity about what cultural sensitivity is and how it should be infused into research and practice. Perhaps nowhere is this more critical than in the area of assessment, with its role in the selection and classification of individuals, the evaluation of treatments, and the testing of scientific hypotheses (Cronbach, 1970). This is also particularly relevant in the United States at this time, given the diversity of cultural groups represented, the radically changing demographic profile, and current social, legal, and political concerns about the disadvantaged status of many ethnic minority groups.

Indeed, issues of cultural sensitivity in assessment have been at the forefront of educational reform over the last few decades. Given the poorer performance on standardized achievement and aptitude measures and disproportionate placement of certain groups of children into special programs, federal mandates such as Public Law 94-142 have been enacted, requiring school districts to ensure nondiscriminatory assessments. Faced with the difficult task of operationalizing such mandates, educators and psychologists have devoted substantial efforts toward reducing bias in standardized testing. From these efforts, a large literature on test validation and test bias has emerged (e.g., Berk, 1982; Cole, 1981; Geisinger, 1992; Messick, 1980; Olmeda, 1981).

On the other hand, developmental psychologists have devoted less systematic efforts to defining a culturally sensitive research agenda and establishing guidelines for culturally sensitive assessments. In fact, researchers often believe that mere inclusion of children from diverse cultural backgrounds is suf-
8. Culture and Assessment

Discussions of cultural sensitivity often occur without prior discussions of what is meant by culture. This may be due, in part, to the fact that there is considerable disagreement among scholars as to the precise meaning of the term. In its broadest sense, culture refers to a way of life of a social group that includes shared norms, beliefs, values, and language, as well as shared organizations and institutions. Clearly, culture is not a unitary phenomenon, but rather a highly complex social context. In order to evaluate the influence of culture on the assessment process, we must begin by specifying those aspects that are most important.

Perhaps the most relevant feature of culture is its role in determining the framework individuals use for making sense of the world—what could be viewed as the social—cognitive component of culture. From this vantage point, culture has been described as a learned system of meaning, transmitted both intra- and inter-generationally, that affords a shared foundation for understanding and interpreting reality (Davis, 1948; Geertz, 1973; Nobles, 1991; Rohner, 1984). This meaning goes beyond language and symbols and provides individuals with a template for organizing social experience, although individuals in a given culture still vary greatly in terms of how they interpret and understand their social world. Such a template includes implicit norms regarding appropriate feelings and behaviors in specific situations, often referred to as the subjective culture of a group (Triandis, 1972), as well as particular values and worldviews, also referred to as the fundamental culture of a group.

By emphasizing how culture provides an organizing framework for the way individuals think and respond, we assume that a child’s cognition, affect, and behavior are influenced by the cultural context of development. That is, one’s active construction of reality occurs in a set of culturally available explanatory models or frameworks for understanding reality (although individuals in multicultural societies may also alter those cultures by integrating different perceptions from mainstream or other cultures). In any case, the validity of an assessment protocol hinges on an awareness of the possible ways individuals in a given culture can make sense of their world. Consider the following interview with Kpelle man described by Scribner (1978):

**DETERMINING AN INDIVIDUAL'S CULTURAL FRAMEWORK**

了解研究参与者的文化框架是相对容易的，当同质社会群体如Kpelle被研究时。人类学家已经对在不考虑种族分类时，被社会群体特征化为高比例的理性主义的框架规范，例如，世俗文化或极其宗教的群体。同样，跨文化心理学研究经常将个体与易于识别的相互影响的文化联系起来。然而，在一个多元文化的多元社会，如美国，个体属于一个复杂社会群体的重叠，包括定义为种族，性别，年龄，宗教，社会阶级，语言，移民身份，少数族裔身份，以及民族识别。确定社会群体或文化，这是最相关的，提供了一个关于文化框架的难点。最常用的方法是根据种族，民族或少数族裔身份，以及民族识别。然而，如我们所指出的，存在实质的同质性在每个群体中，且这些分类方案的易用性。

尽管存在概念的，操作的，和辩解的问题，与种族本身的定义有关，种族主义以持续的特征，随着种族的特性和特定的心理属性。有趣的是，一些非洲中心学者也声明了基因和文化的相互影响。尽管可能存在一些遗传的倾向，被找到的群体的种族，它似乎导致了焦点放在种族上将文化指定给文化只导致了Appiah（1992）所提到的外显种族主义。进一步，给定的这些种族，分类由种族排除了任何类型的微小区分，基于在不同社会群体之间的差异经验，往往根据种族，种族和压迫，稀疏，和混合种族的困难，经常被定义为多元性，社会阶级，语言，移民身份，少数族裔身份，以及民族识别。这些集体身份意味着至少两个群体之间的接触。这些构想的存在导致了心理学的框架研究过程中的一个关键步骤。这种已被认为是文化的验证性，以及它与外部验证性之间的关系已被指出（Washington & McLoyd, 1982）。没有这样的知识，识别相关共内和共外的构想，有效评估这些构想成为问题的。

尽管存在概念的，操作的，和辩解的问题，与种族本身的定义有关，种族主义以持续的特征，随着种族的特性和特定的心理属性。有趣的是，一些非洲中心学者也声明了基因和文化的相互影响。尽管可能存在一些遗传的倾向，被找到的群体的种族，它似乎导致了焦点放在种族上将文化指定给文化只导致了Appiah（1992）所提到的外显种族主义。进一步，给定的这些种族，分类由种族排除了任何类型的微小区分，基于在不同社会群体之间的差异经验，往往根据种族，种族和压迫，稀疏，和混合种族的困难，经常被定义为多元性，社会阶级，语言，移民身份，少数族裔身份，以及民族识别。这些集体身份意味着至少两个群体之间的接触。这些构想的存在导致了心理学的框架研究过程中的一个关键步骤。这种已被认为是文化的验证性，以及它与外部验证性之间的关系已被指出（Washington & McLoyd, 1982）。没有这样的知识，识别相关共内和共外的构想，有效评估这些构想成为问题的。

尽管存在概念的，操作的，和辩解的问题，与种族本身的定义有关，种族主义以持续的特征，随着种族的特性和特定的心理属性。有趣的是，一些非洲中心学者也声明了基因和文化的相互影响。尽管可能存在一些遗传的倾向，被找到的群体的种族，它似乎导致了焦点放在种族上将文化指定给文化只导致了Appiah（1992）所提到的外显种族主义。进一步，给定的这些种族，分类由种族排除了任何类型的微小区分，基于在不同社会群体之间的差异经验，往往根据种族，种族和压迫，稀疏，和混合种族的困难，经常被定义为多元性，社会阶级，语言，移民身份，少数族裔身份，以及民族识别。这些集体身份意味着至少两个群体之间的接触。这些构想的存在导致了心理学的框架研究过程中的一个关键步骤。这种已被认为是文化的验证性，以及它与外部验证性之间的关系已被指出（Washington & McLoyd, 1982）。没有这样的知识，识别相关共内和共外的构想，有效评估这些构想成为问题的。
African Americans must negotiate the Afrocultural, minority, and Anglocultural realms of experience.

The heterogeneity in the African American population also extends to factors such as regional identification (e.g., Southerners versus Northerners) and social class. For instance, although African Americans are disproportionately represented among the lower classes, African American status is not synonymous with being poor, and African Americans are found across the social classes. Scholars often fail to disentangle the African American cultural framework in terms of the influence of both ethnic identity and social class. Yet, it is likely that such differences would impact an individual’s cultural framework. For instance an upper-middle-class, African American man might have a cultural framework that appears more Eurocentric given the European nature of many middle-class artifacts. In contrast, a lower-class, African American male whose artifacts are more likely to be identified as Afrocentric by both Whites and Blacks should have a more Afrocultural perspective.

Similar issues arise with the classification of individuals as either Hispanic or Latino. Although the terms have been used to refer to individuals with a common linguistic (Spanish) and historical (colonized by Spain) heritage, there is so much variation within the Latino population in terms of country of origin, ethnic identity in country of origin, acculturation status, and demographic characteristics that the utility of such overly general labels is limited. Although the Triple Quandary model has not been applied to Latinos, it seems useful because Latinos must also negotiate three realms of experience—Latino, minority, and Anglocultural. However, although most African Americans have lived in the United States for multiple generations and share a common language, for Latinos, their country of origin, generational history, and language preference varies greatly and must be considered simultaneously.

Although ethnicity can serve as a shorthand for classifying individuals for evaluation, it is important to bear in mind that it should be considered a first step in evaluating an individual’s cultural framework, and several potential serious limitations must be acknowledged. Many of these limitations are operational wherein common methods of classification can lead to undesirable errors. For example, classification schemes virtually fall apart when considering racially or ethnically mixed children or children living in adoptive families. Similarly, relying on surnames for identification of Latino subjects obscures such dimensions as generational status so that a fifth-generation Latino child would probably lead to the conclusion that individuals in Culture B are deficient. This would be particularly true if the assessment were conducted by researchers from Culture A without inclusion of researchers from Culture B. Such normativistic approaches have dominated the field of developmental psychology in the United States, promoting comparison of children from other cultures with White, middle-class standards (Azibo, 1992; McLoyd & Rendolph, 1984). Not only has it been presumed that such Eurocentric standards are superior, but they regularly have been embraced as the most appropriate evaluative referent for people of all cultural groups. As a consequence, failure to match certain Eurocentric ideals has often led to the conclusion that certain groups of children are inferior, deviant, or abnormal.

In developing fair assessments for members of distinct cultural groups in developmental studies, an overarching issue is whether the constructs themselves are free from bias. That is, if a construct that is meaningful in Culture A is meaningless in Culture B, assessment of this construct in Culture B would probably lead to the conclusion that individuals in Culture B are deficient. This would be particularly true if the assessment were conducted by researchers from Culture A without inclusion of researchers from Culture B. Such normativistic approaches have dominated the field of developmental psychology in the United States, promoting comparison of children from other cultures with White, middle-class standards (Azibo, 1992; McLoyd & Rendolph, 1984). Not only has it been presumed that such Eurocentric standards are superior, but they regularly have been embraced as the most appropriate evaluative referent for people of all cultural groups. As a consequence, failure to match certain Eurocentric ideals has often led to the conclusion that certain groups of children are inferior, deviant, or abnormal.

Reflected in these normativistic models are a set of core Eurocentric values including effort optimism, material well-being, possessive individualism, egalitarian-based conformity, the democratization of equality, and a person-object orientation (Baldwin, 1991; Boykin, 1983). Indeed, several influential constructs in developmental psychology including internal locus of control, delay of gratification, and the notion of an independent self system seem to flow directly from these core Anglocultural values. Moreover, these core values are grounded in an assumption of the psychic unity of humankind that both minimizes attention to the sociocultural context of development and leads to a denial of the cultural influences inherent in Anglo-American models of behavior.

The cognitive-developmental literature on children’s moral reasoning provides a case in point. Perhaps the most controversial aspect of the work of
Kohlberg and his colleagues (Kohlberg, 1984; Kohlberg, Boyd, & Levine, 1990) has been the claim that the development of moral reasoning follows an invariant sequence toward the same justice-based universal principles regardless of cultural settings. Much of the debate surrounding this assertion stems from the robust finding that certain groups of people (e.g., women and individuals from less industrialized societies) are less likely to utilize higher stages of moral reasoning. In fact, in some societies, the higher stages of post-conventional reasoning are markedly absent (for a review, see Snarey & Keljo, 1991). One interpretation of this finding is that certain societies do not promote the development of more advanced stages of reasoning. However, it is also possible that Kohlberg's stage model represents an individualistic, elitist, Western theory that fails to account for the cultural perspectives of other groups, and that the assessment methodology obscures those perspectives, as many critics have claimed (Gilligan, 1982).

Thus, the concept of justice on which Kohlberg's theory is based simply may be less relevant for certain cultural groups, and assessments of moral reasoning using justice criterion may lead to conclusions that certain individuals or groups of individuals are less advanced. In this manner, it is plausible that the basis for moral thought may vary significantly across cultures, and that a lack of understanding of such variation may lead to premature conclusions about the moral development of certain groups. Although Kohlberg defended his notion of a justice-based morality on a priori philosophical grounds consistent with his Kantian perspective, it is clear that his model does not adequately fit the diversity of cross-cultural perspectives on moral principles. By using narrowly defined models of moral reasoning to assess and measure moral development, cultural variability in moral discourse and ways of understanding is often suppressed.

Thus, the psychological enterprise is presumed to be objective and value neutral, although, in fact it is culturally loaded in favor of the Westernized ideals of White, middle-class U.S. culture, both in determining what constructs should be studied and in specifying how they should be operationalized. It has been proposed that such cultural imperialism has been useful for those who wished to justify the domination and disenfranchisement of people from selected cultural groups, and several scholars have pointed to the instrumental role of western social science in protecting the cultural agenda of Anglo-Americans (Sampson, 1977; Stanfield, 1985).

Not only have nondominant groups been described as disadvantaged or deviant, but relatively little is known about the development of children from diverse cultural backgrounds (McLoyd & Randolph, 1984; Ogbu, 1981). It is important to recognize that members of ethnic minority groups are part of the broad diversity of cultures in the United States that merit attention because of the unique value of their cultural contributions. In addition, this knowledge base is important for cross-cultural comparisons and for generating and testing theories of universal principles of behavior. However, the development of such a knowledge base hinges on the development of assessments that accurately reflect the variables of interest as expressed in distinct cultural groups. From this perspective, the resolution of the familiar emic-etic debate must be seen as an ongoing process. That is, assessments must be developed that enable researchers to assess both commonalities and differences across cultures.

Decision making in research must also coincide with the practical applications of the data. In some cases, it may be critical to examine developmental processes in a particular cultural group at a specific point in time, or even in a particular subgroup in that culture. For example, if high-school dropout rates were determined to be exceedingly high among first-generation Mexican children of migrant farmworkers in California, it would be valid to examine the specific cultural factors and relevant constructs in that population, relying heavily on ethnographic methods. Of course, it is likely that these findings would not generalize to other populations and would have only limited value in furthering our understanding of the causes of school dropout. In other cases, practical demands may dictate the search for universals. For example, a number of social institutions (e.g., public schools) must provide relatively uniform services to all students based on some set common standards. Unfortunately, the search for universal principles have been restricted to largely white, middle-class children and must be reworked to include children from diverse cultural groups.

**INSTRUMENTATION AND METHODOLOGICAL THREATS TO VALIDITY**

At the forefront of issues plaguing developmental researchers are concerns about the validity of assessments and potential systematic biases based on membership in a particular cultural group. In this section, we examine these concerns, focusing primarily on criterion, content and construct validity and related biases, as well as problems stemming from data collection methods and procedures. Because of the wide range of variables measured, methods employed, and assessment formats used in developmental research, we cannot review each separately but can only provide a general overview of the major concerns.

Both validity and bias refer to a collection of concepts involving the ability to make accurate inferences based on test scores. An inference from a test score is presumed to be valid when a variety of differently types of evidence support its plausibility. Bias has been defined as the differential validity of a test score for any definable group of subgroup of individuals (Cole & Moss, 1989). A complete review of the psychometric literature of validity and bias as
related to the assessment of minority children and youths is beyond the scope of the chapter and can be found elsewhere (e.g., Berk, 1982).

In general, the study of test validation and particularly test bias has been dominated by concerns over criterion validity, that is, the differential accuracy of predictions for members of different cultural groups. Given the common practice of using assessments for selection, there has been much concern that tests that predict the performance or behavior of White, middle-class children may not accurately predict the performance of ethnic minority children, resulting in a selection bias. To examine whether selection bias has occurred, the relation between tests scores and relevant criteria are modeled by regression lines and compared across groups, with significantly different regression lines across groups indicating selection bias. Interestingly, most well-controlled studies of selection bias have found that the majority of standardized tests are equally valid for members of ethnic minority groups. In fact, if anything, use of overall regression lines leads to overprediction rather than underprediction of the criterion performance of minorities (Cole, 1981; Hunter, Schmidt & Hunter, 1979). Of course, it is also important to remember that the ultimate test of criterion validity is the validity of the criterion itself.

Content validity refers to the extent to which a particular sample of items is reflective and representative of the content of the domain of interest. Content bias occurs when the items on a test are less familiar to certain groups of people or when cultural norms promote patterns of responding that may differentially influence responses. It is probably the most well known and easily understood type of bias, although it may occur for a number of different reasons.

Perhaps the most common type of content bias is due to differences in language and meaning. Although concerns about language proficiency literally beg the question because common sense alone would point to the need to test people in their primary language, there has been remarkable inconsistency in actual practice. For example, Busch-Rossnagel (1993) reviewed developmental studies on Hispanic children. To examine practices with regard to the language of assessment, she examined seven empirical studies conducted on Latino children that were published in a special (1990) Child Development edition on ethnic minority children. She concluded:

No study in the Child Development special issue reported the percentage of subjects tested in English versus Spanish or tested for a language-of-testing effect. . . . Three studies, all using school children, included measures available only in English. Of the four studies with Spanish measure, two reported using the method of back translation, and two did not report how the Spanish versions were developed. Only one of the studies took the necessary step of pretesting the measures with the non-Anglo population to look at their validity, and one study simultaneously developed English and Spanish versions of the measures. (p. 199)

The testing of children and youth in their primary language and the utiliza-

tion of translations that are equivalent in meaning are obvious procedures in the assessment of non-English-speaking children. A common technique for establishing equivalence in meaning has been the back-translation method. This is accomplished by having two bilingual individuals translate the measure—first one person translates from the source language to the target language, then the second person back-translates from the target language to the source language. Comparisons in the translations are made and adaptions are made to optimize comparability.

This procedure is preferred over standard translations, although problems can still occur if bilingual translators do not share the culture and status of the respondents. For instance, two upper-class Puerto Ricans might translate and back-translate a measure from English to Spanish that would nevertheless be difficult for a lower-class Mexican farmworker to understand. Clearly, the relevance of the words and phrasing must be established for the respondent group. This issue is not limited to linguistic minorities but is applicable in any group where the meaning of language may vary systematically. For instance, in contemporary youth culture in the United States word meanings are commonly reversed, such that bad means good, and cool means uncool, and so forth. In this case, some type of preliminary evaluation of the wording of the measures must be established.

Another related area that is frequently overlooked involves the relevance of the specific items and item format for the cultural group studied. Questions and items developed on White, middle-class youths raised in the United States often reflect a content domain specific to that cultural group. Even if a child understands the language, the meaning may still be irrelevant. For instance, measures of family functioning may include referents to largely middle-class luxuries such as owning one's own home and taking family vacations. Similarly, pictorial materials often show only White children or families engaging in typically White, middle-class activities.

Culturally defined response styles may also limit the validity of certain types of assessments. Response style refers to a characteristic way of responding that is not related directly to the content of the assessment but to how items are worded. For instance, "si-ismo" in the Latino culture refers to the tendency to say "yes" or "si" to indicate agreement with others (particularly authority figures), independent of one's actual ideas. In fact, the tendency to agree with questions regardless of content, often referred to as ya-seeing, is one of the more widely studied biases (Shuman & Presser, 1981). In some cases, this can be controlled for by using balanced scales with equal numbers of positively and negatively worded items. However, this is sometimes problematic with younger children, particularly when double negatives result (e.g., "Other children don't like me"—Yes or No).

Some studies have found differences in the tendency of children from certain cultures to favor extreme responses. For example, Bachman and O'Mal-
factor analysis with the Mexican sample failed to confirm the original factor structure and resulted in a virtually uninterpretable conglomeration of factors. The authors concluded that the translations were not sensitive to the differences in construct meanings across cultures. They developed a new instrument based on culturally relevant definitions of the three original constructs. With this approach, factor analysis revealed three factors that were roughly comparable to the original definitions.

Aside from concerns about the validity of the assessment instruments, there are additional concerns related to assessment procedures and methods that must be considered when working with minority children and youth. These stem from differential perceptions of the research/assessment enterprise by minority children and families and include low response rates, problems in eliciting responses, and response inaccuracies. For instance, members of disadvantaged minority groups may distrust academic (particularly non-minority) researchers and suspect that findings will be used to perpetuate their disadvantaged or inferior status. In immigrant groups, one's immigration status as an illegal alien may preclude identification or involvement in any formal scientific investigation. Low response rates can severely limit the representativeness of an initial sample. Furthermore, differential attrition of minority participants is often found. Even those who do participate in the interview may be reluctant to take it seriously or respond in a socially desirable manner, and mechanisms for eliciting responses (e.g., direct questioning versus open-ended discussions) may differ significantly across cultures.

Such concerns are partially remedied by conducting extensive pretesting of both instruments and procedures in the culture to be studied, including the use of traditional ethnographic methods such as participant observation and interviews with local informants. Focus groups are frequently used to shortcut this process and provide general guidance for the assessment. This process can also shed light on some of the culturally important questions to ask that may have been neglected in the original research formulation. Finally, the use of indigenous interviewers and experimenters throughout the research process can minimize data collection problems.

CONCLUSION

Developmental researchers must work toward the development and utilization of culturally sensitive assessments with children and youths from diverse cultural groups. As we have discussed, this is a complex task that must be infused into the research process on multiple levels. First, given the normativistic models based on White, middle-class children that have dominated developmental psychology, cross-cultural studies must be conducted to determine whether traits, characteristics, constructs, and models presumed to be univer-