# CHAPTER 6

# Fairness

Test construction experts are beginning to make a distinction between test fairness and test bias. Traditionally, a test was considered fair, or fair(er), based on evidence showing the extent to which it was free from *test bias*; and, as bias decreases so does construct-irrelevant variance within the score. When construct-irrelevant variance is reduced, so are systematically different levels or patterns of performance as a function of membership across various groups, assuming the groups possess about equal levels of the ability under consideration. Importantly, test authors now realize that a test may be free of bias in a technical sense but may not be *fair* for a particular examinee. That is, scores may be influenced by cultural or linguistic characteristics that can impact performance negatively if the test contains culturally loaded content; alternatively, the test may be biased, but not culturally loaded.

From a psychometric perspective, *bias* is a technical term, operationalized by results from particular statistical tests (e.g., correlational and related factor analytic and model-testing comparisons across groups). For example, a test that is biased against a particular group may predict less well for the marginalized group, may yield a factor structure for the marginalized group that is different from the mainstream group, and, in general, may produce more error in the scores or the prediction for the marginalized group relative to the mainstream group. Importantly, even though a test may be free of technical bias, examiners may still want to minimize the effects of culture or language influences. The underlying assumption is that culturally- and linguistically-loaded tests may be *unfair* for some examinees (i.e., those who are unfamiliar with the particular culture and/or who do not possess the skills to use the language embedded in the tests). Consequently, examinees will want to consider the extent to which particular tests/subtests contain culturally or linguistically loaded content before they use them for certain examinees. Fortunately, the field has advanced sufficiently that test authors can evaluate a test to determine the influence of technical bias *and/or* cultural/linguistic characteristics and reduce both. In this section, we describe some of the strategies embedded in the conceptualization and design of UTAGS to optimize fairness and minimize bias for individuals varying in age, sex, race, ethnicity, language, and nationality.

Demonstration of fairness requires examination of the *internal* characteristics of a test as well as the test's relationship to relevant *external* variables. The extent to which the UTAGS is free from bias and fair is based on theoretical, statistical, and practical aspects of its development and application. In this section, we discuss multiple lines of evidence to support the conclusion that UTAGS provides a fair operationalization of the constructs it assesses. We begin with a discussion of the theoretical rationale for the test structure.

# TEST THEORY AND FAIRNESS

The UTAGS was formulated with an underlying model of fairness positing five core concepts:

1. UTAGS's directions encourage examiners/raters to ignore language-related differences/limitations of examinees and to focus on performance in the six areas of cognition, creativity, leadership, literacy, math, and science (e.g., an examinee's ability to "explain complicated concepts effectively," a cognitive item, may be expressed in drawings, gestures, and/or pantomime, as opposed to verbalizations);
2. UTAGS's ability to assess multiple elements of ability is fairer than one that assesses fewer or a single dimension of ability;
3. UTAGS provides assessment of a significant number of novel abilities in the nonacademic areas, particularly the Creativity subtest, which minimizes to some extent the impact of previously acquired knowledge ability and provides a fair(er) assessment relative to those instruments that rely on crystallized abilities;
4. UTAGS's items are written to minimize speeded performance, and, consequently, the test is fairer than instruments that emphasize speed; and
5. UTAGS allows examiners/raters to compare examinees to local peers. Each of these concepts has a rich theoretical and applied background, which is summarized in the following sections.

UTAGS was formulated to reduce the bias problems inherent in heavily language-dependent tests. Individuals who are not native speakers of English may be disadvantaged by English-language procedures (e.g., Duran, 1989; Geisinger, 1992; McCallum, 2013; Oakland & Parmelee, 1985); also, examinees who are speech- and language-impaired may be disadvantaged by testing procedures in *any* language (e.g., Braden, 1994). Because of the wording of the directions, examiners are less likely to penalize examinees who exhibit speech impediments, unusual dialects, and who come from other countries/cultures. Of course, even tasks that require no receptive or expressive language still contain elements of the culture in which they are developed and are not entirely language-free, and UTAGS authors recognize that no test is culture-free.

Second, the UTAGS includes indexes of multiple abilities, some of which are less culturally determined than others (e.g., creativity, leadership). For example, items within those subtests require the rater to assess the examinee's ability to produce creative drawings, improvise, display an active imagination, demonstrate flexibility, show respect, and so on. These characteristics are considered valuable across most cultures/ethnicities/races.

Third, UTAGS was designed to provide some assessment of abilities, particularly fluid reasoning and problem solving, that are less influenced by educational opportunity and that are thereby fairer than tasks reliant on previously acquired knowledge. Although UTAGS does provide assessment of academic skills (i.e., literacy, math, and science), the nonacademic subtests tend to assess more general, although important, school-based outcomes.

Fourth, the UTAGS places comparatively little emphasis on time as part of the examinee's score (i.e., items rarely reference speeded performance). Timed testing has been implicated as a potential source of testing bias, although its biased properties tend to be small (Jensen, 1980). Decreased reliance on timed performance has been recognized as important (e.g., Wechsler, 1991), principally because of the emphasis on accuracy over speed in intelligent behavior. Pure speeded tasks have items that

are generally so easy that nearly all examinees would obtain a perfect score if allowed enough time; however, speeded tasks place individuals with certain exceptionalities (e.g., motor impairments) or cultural backgrounds at an unfair disadvantage (Knapp, 1960).

Finally, the UTAGS provides examiners/raters the opportunity to rate examinees compared to local peers. Behavior that is considered "average" in Manhattan, KS, may not be average in Manhattan, NY. As we discuss in considerable detail in the Local Norming chapter, examiners have two choices for engaging local norming strategies: the raw score, or Rational method, and the local standardization, or Empirical method. Of course, national norms are provided for those examiners who prefer to compare examinees' performance to national peers. Below we discuss some of the statistical evidence addressing UTAGS's fairness.

## INTERNAL TEST CHARACTERISTICS

The internal characteristics of a test include the content, statistical properties, and structure of its items, subtests, and composite scales. The extent to which internal characteristics are fair can be evaluated by multiple methods, including expert reviews of item, subtest, and scale content; statistical analyses of item, subtest, and scale fairness; comparative analyses of group mean performance by sex, race, ethnicity, and language; and comparison of factorial structure across groups.

**Test Content and Procedures.** UTAGS was designed to be as fair as possible in terms of item selection, item composition, and examiner characteristics. Items were chosen to minimize cultural, ethnic, and racial influences. The abilities assessed by the items are considered valuable by all educators and parents. In addition, there is no a priori reason to believe that the items will interact with examiner/rater characteristics to favor examinees from any particular background. UTAGS test instructions were created to be equally intelligible and understandable by all examiners, independent of their group membership, and are brief, concise, and clear. Of course, we encourage examinees to be particularly vigilant in addressing potential sources of biases in their ratings.

**Expert Review of Item and Subtest Content.** UTAGS items and procedures were reviewed for potential bias during development. Reviews were initially conducted by the authors and the test development staff. During two additional phases of development (i.e., during item pilot/tryout studies), psychologists and consultants representing diverse cultural, ethnic, and racial backgrounds also reviewed the content of items and procedures for bias. These bias consultants represented the perspectives of female and male respondents; African Americans, Asian Americans, Hispanic Americans, and Native Americans; and do not change as a function of culture. As authors, we consulted with experts in the field, either directly and personally, or by reading the literature, and the UTAGS development was influenced by these experts (e.g., Bracken, 1987; Braden & Anthansiou, 2005; Flanagan, Ortiz, & Alfonso, 2007, 2013; Jensen, 1980; Reynolds & Lowe, 2009),

## EMPIRICAL EVIDENCE OF FAIRNESS

For the UTAGS, a number of statistical analyses were conducted to ensure that items functioned similarly across sex, race, ethnicity, and language. Results from various analyses are reported in the

Normative Information, Reliability, and Validity chapters and will not be repeated here. However, we mention below some of the most salient results from the statistical analyses, show tables that describe the most relevant fairness-related results from various studies, and summarize the results.

Several lines of research address subtest and scale fairness, including evidence related to reliability and validity. For example, see: (a) reliability data across groups demonstrating similar measurement precision for varying groups categorized by sex, race, and ethnicity; (b) construct validity data showing strong evidence in support of the test's six-subtest structure and data showing relations between UTAGS and related measures; and (c) comparison of actual levels of performance between groups with socioeconomic status and demographic characteristics.

## Reliability

UTAGS was designed to have comparable levels of score reliability and precision for all groups. However, without verification of comparable measurement accuracy, the assumption cannot be made that the reliability coefficients of any test scores for a normative population are the same as the reliability coefficients of that test's scores for diverse groups. Accordingly, reliability coefficients were calculated separately for male and female examinees; African Americans, Asian/Pacific Islanders, and Hispanics; and for students identified as gifted and those with learning disabilities. As is apparent from Table 7, coefficient alphas are practically identical for all of these groups; in fact, Alpha for the General Aptitude Composite is exactly the same across all groups, .99. Results indicate that the UTAGS subtests and scales are consistently reliable across sex, race, and ethnicity and that all reliabilities meet Bracken's (1987) standards.

## Construct Validity

Just as measurement precision must be comparable across groups, so must the construct validity of the test. The UTAGS's construct validity is strong, as demonstrated by exploratory and confirmatory factor analyses (see Tables 18 and 19 and Figure 3.) Discussion of the factor analytic results is presented in Chapter 5 and is not repeated here; suffice it to say the data are strong, and examiners can have confidence in interpretation based on the six subtests score as somewhat independent operationalizations of the abilities they are intended to assess, as well as the total score (General Aptitude Index). In addition, UTAGS scores are related significantly to other measures created to assess similar or related constructs (e.g., the *Gifted Rating Scales*, end-of-year standard scores) based on the data in Tables 12 and 13. These data show correlations, coefficients, and mean difference comparisons. Finally, UTAGS scores can be used to predict important real-world outcomes, as indicated by the values in Tables 12 and 13, Figure 3, and the data obtained from the binary classifications within the Receiver Operating Characteristic/Area Under the Curve (ROC/AUC) analyses. These data describe sensitivity/specificity indices at various levels of performance. For example, based on these analyses, the authors suggest a cut score of 110 for screening-for-giftedness purposes. Of course higher cut scores will be selected by educators for placement services, depending on the resources and philosophy of the school system decision makers.

### Group Comparison Studies

Because some test users consider mean score differences between groups an index of fairness, mean difference analyses were conducted for the UTAGS. The underlying assumption is that groups should show equal ability, and if they do not, the test is biased against the group (or groups) obtaining the lower score(s). Others do not consider mean differences de facto evidence of bias. For example, Jensen (1980) referred to the belief that unequal mean scores automatically indicate bias as the "egalitarian fallacy," which is "the gratuitous assumption that all human populations are essentially identical or equal in whatever trait or ability the tests purport to measure" (p. 370). He noted that there is no a priori reason to expect groups who differ along sex, race, or ethnicity to exhibit the same mean score on intelligence tests (or a host of other human variables, such as height, weight, muscle mass, gregariousness, etc.). In fact, Suzuki and Valencia (1997) concluded that racial/ethnic IQ differences are among the most thoroughly documented findings in psychology. Suzuki and Valencia addressed the complexities associated with understanding racial/ethnic differences and concluded that focusing primarily on group differences can lead to misconceptions, particularly in view of the fact that within-group differences exceed between-group differences. Obviously, the use of mean-score differences as an index of fairness is controversial. However, many test users are interested in examining such differences for good educationally relevant reasons. For example, students previously identified as gifted should differ from those who were not. Similarly, students previously identified with learning/cognitive limitations should score lower than those who were not. Data for a number of (group) comparisons were obtained for UTAGS and are presented in Table 17. From scrutiny of these data, it is apparent that there are no mean differences between males and females and between White students and Hispanic students in the standardization sample. Students within these groups score at the population mean or nearly so on all subtests and the GAI (i.e., around 100). There are small differences between White and African American/Black students (5 points on the GAI) and between White students and American Indian/Eskimo students (7 points), with White students scoring slightly higher; however, these mean differences are less than those obtained between White students and minority group students on language-loaded tests. Those differences typically range from 8 to 15 points. Finally, the mean scores between students identified as gifted are typically more than one to 1.5 standard deviations above the population mean of 100; students with learning disabilities and intellectual deficits tend to score about .67 and 2.0 standard deviations below the population mean, respectively, as expected.

Finally, it is recommended that examiners interpret mean-score differences with caution and with the understanding that many influences interact to produce those differences, primarily socioeconomic status (SES). However, these differences can be helpful to examiners as they contemplate the utility of UTAGS for particular students.

## CONSIDERATIONS FOR CULTURALLY/ LINGUISTICALLY FAIR(ER) ASSESSMENT

UTAGS authors attempted to limit culture and the related language confound as much as possible yet still retain strong item discrimination indices. Of course, some would argue that certain measures (e.g., intelligence) cannot and should not be completely devoid of cultural influences if the intent is

to predict performance within that particular culture (Reynolds & Lowe, 2009). Nonetheless, experts in the field have contributed significantly to the creation of guidelines that can inform practitioners about culture and language loadings, and hence help examiners assess and/or reduce the influence of these variables to a considerable extent. For example, according to a model of nondiscriminatory testing (Ortiz, 2002), examiners should: (a) develop culturally and linguistically-based hypotheses; (b) assess language history, development, and proficiency; (c) assess effects of cultural and linguistic differences; (d) assess environmental and community factors; (e) evaluate, revise, and retest hypotheses; (f) determine appropriate languages of assessment; (g) reduce bias in traditional practices; (h) use authentic and alternative assessment practices; (i) apply cultural-linguistic context to all data; and (j) link assessment to intervention. Although testing is only a small part of this overall model, use of tests that minimize some cultural/linguistics influences, as the UTAGS does, can reduce bias and increase fairness in the assessment process.

## CONCLUSIONS

A major goal for the development of the UTAGS was to ensure fairness through the use of multiple methods and analyses. In this chapter, these methods were discussed, including expert bias reviews of procedures and item content, and extensive analyses of reliability and validity data (e.g., analyses of the comparability of the UTAGS's measurement precision, factor structure, and numerous group comparison studies). Data support the use of UTAGS as a fair(er) instrument for screening students who may be talented or gifted, in part because of the local norming methods available for scoring and the directions to raters to ignore atypical communication strategies of examinees.

The data provided in the UTAGS manual should be sufficient to establish the UTAGS as a useful addition to current methods of identifying and screening students for special needs, at both extremes of school success (e.g., those with intellectual and learning disabilities and those who are gifted). The UTAGS format requires that an examinee's behavior be compared to that of his or her peers as rated by a qualified individual—usually the student's teacher—and benefits from the capacity to make local comparisons. Results presented in this manual help establish validity of the instrument. Data strongly support use of UTAGS for the purpose of identifying students who are at-risk for school failure and those who are high-functioning/gifted.