A Primer for Evaluating Test Bias and Test Fairness: Implications for Multicultural Assessment

Richard S. Balkin, Courtney C. C. Heard, ShinHwa Lee, and Lisa A. Wines
Texas A&M University-Corpus Christi

The authors present a model for evaluating test bias and test fairness in the practice of assessment in counseling. An explanation of steps evaluating the appropriateness of instruments related to cross-cultural comparisons is presented, as well as issues of use and misuse of test scores. Implications to counselors, and specifically to professional school counselors, are highlighted.

Issues of test bias and test fairness are widely known points of concern among counseling professionals. Test bias occurs when a group or several groups experience differences in scores on a test or varying interpretations based on similar tests scores as other groups (Balkin & Juhnke, 2014). However, simply because various groups perform differently on a test, thereby indicating test bias, does not mean that the aforementioned test is unfair. For a test to be unfair, the bias between or among test scores for groups should be supported with a viable theoretical framework (Balkin & Juhnke, 2014). Examples of deviations from test fairness occur when (a) the uses of scores are not utilized and interpreted the same across all participants, such as having different cut scores across a variety of demographic factors; (b) the opportunity to prepare and/or complete the instruments is not the same for all participants, such as standardized instructions, tasks, and preparation; and (c) the conditions in which the test is administered is not uniform, such as variations in test environment.

Author Note:

Richard S. Balkin, Courtney Heard, ShinHwa Lee, and Lisa A. Wines, Department of Counseling and Educational Psychology, Texas A&M University-Corpus Christi.

Correspondence concerning this article should be addressed to Richard S. Balkin, Texas A&M University-Corpus Christi, Counseling and Educational Psychology Department, College of Education, ECDC 232, 6300 Ocean Drive, Unit 5834, Corpus Christi, TX 78412-5834. Email: richard.balkin@tamucc.edu

Test fairness and bias can have a deleterious effect on clients and students who seek or benefit, directly or indirectly, from counseling services across the various spectrums (e.g., clinical, community/agency, school-based, college, rehabilitation, etc.). Relevant to the discussion of test bias and test fairness are the procedures in place that are used to protect and advocate for clients and students. The examination of factor invariance is one such procedure. Testing for *factor invariance* refers to the process of evaluating evidence that the properties and interpretations of test scores are similar across various groups (Dimitrov, 2011). For example, many psychometric instruments were normed with predominately White samples. Evaluating invariance among different ethnic groups may be appropriate to address the extent to which the factor structure of an instrument is consistent among scores of various ethnic minorities. Factor invariance procedures involve utilization of latent variable modeling (LVM; e.g., confirmatory factor analysis [CFA]) to ascertain whether or not scores from different groups demonstrate the same factor structure on an instrument. While a description of LVM is outside the scope of this manuscript, LVM is statistically more sophisticated and requires more advanced software (e.g., LISREL, AMOS, Mplus). A more simplistic method would be to evaluate the factor structure using exploratory factor analysis (EFA) for various groups. However, such analyses are dependent upon sufficient sample sizes to conduct factor analytic procedures (whether EFA or CFA). Many instruments do not have sufficient samples from minority populations. Thus, testing for factor invariance is rarely demonstrated in test manuals or conducted initially by developers. Rather, researchers in counseling and education need to conduct independent studies to evaluate factor invariance on specific tests.

Although counselors may view differences in scores based on ethnicity to be a variable of interest, if instruments created to measure theoretically tenable constructs contain scores that vary across ethnic backgrounds, such findings may actually be inconsequential. In this case, a *construct* is a phenomenon that cannot be directly observed (e.g., mood, affect, personality, intelligence, achievement, aptitude, interests) but can be measured through the development of assessment instruments. The measurement of a construct is dependent upon an operational definition that is supported through theory. *Construct irrelevant variance* refers to the "extent to which test scores are influenced by factors that are irrelevant to the construct that the test is intended to measure" (AERA, APA, & NCME, 1999, pp. 173-174). Such differences are a subject of interest in many types of testing. We provide two heuristic examples centered on construct irrelevant variance.

Testing between ethnic groups on achievement test scores is common practice in educational settings, and such findings substantiate evidence of an achievement gap among many ethnic groups compared to White students. However, the very nature of testing between groups implies a relationship between the groups (i.e., the independent variable) and achievement test scores (i.e., the dependent variable; Thompson, 2006). Assessment professionals in counseling and education should proceed cautiously, as the underlining notion that ethnicity is related to academic achievement is offensive and serves as the crux of this heuristic example on construct irrelevant variance. In terms of identifying an operational definition of achievement and variables related to achievement, ethnicity is not a factor. Therefore, the postulation that ethnic differences should be analyzed in academic achievement suggests that ethnicity is a viable variable in such an evaluation.

For example, assessments may be thought to be unfair and in favor of the middle-class Caucasians, who may often have access to resources that may bolster improved performance. The American Counseling Association Code of Ethics (2005) addressed that counselors operate with cultural sensitivity when choosing an assessment for use with diverse populations, as well as administering and interpreting the results obtained. However, many of the assessments utilized in diagnosis of behavioral and cognitive deficits, substance abuse issues, and mental disorders continue to be normed using samples underrepresented by ethnic minorities and other diverse groups. Is the underrepresentation of diverse groups in a normed sample indicative of an unfair assessment? Most achievement assessments widely used have been shown to be valid and reliable measures of the construct, making it a fair assessment of the individual's capabilities. The inclusion of the variable ethnicity in contributing to academic achievement may influence the fallacy of these studies. This provides implications that group differences in scores are based on ethnic group membership, which may further perpetuate discriminatory, racist, and prejudicial attitudes towards ethnic or other minority groups.

Ethnicity was used to explain differences in scores on assessments used for attention and behavioral problems, mental health diagnoses, achievement, and intelligence (Morley, 2010; Rabiner, Murray, Schmid, & Malone, 2004; Whaley, 2004). Rabiner, Murray, Schmid, and Malone (2004) explored the relationship between ethnicity, attention problems, and academic achievement in a sample of Caucasian, African American, and Hispanic first graders. The authors reported that being African American was a significant positive predictor of inattention. In addition, nearly half of the achievement gap between African American and Caucasian students in the sample were associated with ethnic group differences in problems with attention (Rabiner et al., 2004, p. 503). The results appear to indicate that one's ethnic group membership may contribute to inattention, a factor that may influence achievement. The implication of these results is that the achievement differences between Caucasian and African American students may be influenced by attention problems, which are more prevalent among African American children. Ethnicity is not included in the definition of inattention or achievement, thereby making it an irrelevant contribution to understanding group differences among these constructs. Elwy, Ranganathan, and Eisen (2008) conducted a study assessing race/ethnicity and diagnosis as predictors of outpatient service utilization among clients initiating treatment. The results of the study indicated that Latinos and Blacks, as compared to Whites, reported greater symptom and problem frequency and/or severity related to comorbid mental health and substance abuse problems. However, there was not a statistically significant relationship between racial-ethnic group membership and the number of outpatient visits. The authors failed to provide justification for their use of both the terms race and ethnicity in relation to mental health and substance abuse symptom severity and frequency. Race and ethnicity are two separate terms with different definitions, yet Elwy et al. seemed to utilize them as one. As demographic variables, race and ethnicity are not underlining psychological constructs, yet they are treated as such in many research studies (Beutler, Brown, Crothers, Booker, & Seabrook, 1996). Beutler et al. (1996) stated, "[A]dhering to unsubstantiated assumptions of the immutability of demographic descriptors could work either to further enfranchise or to disenfranchise existing social, economic, and political power structures" (p. 892).

While authors of extant research present group differences based upon ethnicity, such illustrations are reprehensible due to reasons mentioned above. In addition, counselors need to be aware that not all instruments measure constructs adequately across various multicultural groups. Another heuristic example using the Career Search Efficacy Scale (CSES) follows.

The CSES was developed by Solberg et al. (1994). Seventy-two items were initially developed related to three domains: career exploration, job exploration, and personal exploration. Solberg et al. conducted a principal component analysis (PCA) on the original 72 items using scores form 192 college students, predominately female (n = 110, 57%) and White (n = 168, 87.5%). The PCA resulted in four identified components, accounting for 67.6% of the variance in the model: job search efficacy, interviewing efficacy, networking efficacy, and personal exploration efficacy. The emerging structure was different from the hypothesized structure. Additional limitations include conducting a PCA instead of an EFA (Dimitrov, 2011) and using a relatively small data set with an initial set of 72 items. Stevens (2009) recommended that five to ten participants per item, though samples of 300 to 500 participants tend to be relatively stable. Moreover, Nota, Ferrari, Solberg, and Soresi (2007) attempted to validate and adapt the CSES with Italian youth and found a three-factor solution accounting for 48% of the variance in the model. Thus, limitations in the initial validation sample by Solberg et al. (1994) may produce confounding results as evidenced by variability in the factor structure from a separate and distinct cultural group. In this case, the culture of the participants does appear to have an effect on the measure of the construct, career search efficacy.

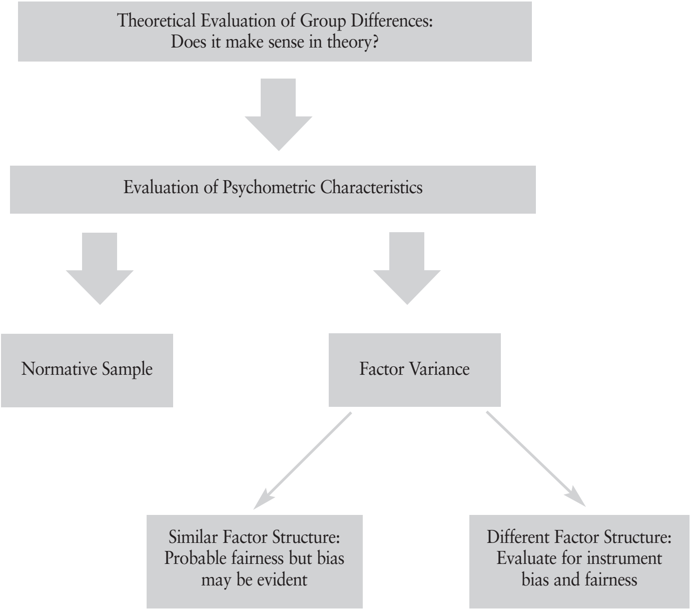
A Model for Evaluating Test Fairness

Counselors should be informed about social and cultural factors that impact administration, scoring, and interpretation of assessment instruments (CACREP, 2009, section II.7.f). A model focusing on the evaluation of test fairness and bias may be helpful in informing counselors about using assessment instruments in practice, particularly with diverse populations, and evaluating counseling research, which often include assessment instruments in the study. Figure 1 shows a visual model of this process.

Evaluating theory. An essential component to establishing evidence of test validity is the demonstration of evidence of test content (AERA et al., 1999). Establishing a connection between the items and extant theory and literature, along with expert reviews of items, represents typical procedures

Figure 1.

Model of Evaluating Test Bias and Test Fairness


Theoretical Evaluation of Group Differences:
Does it make sense in theory?
Evaluation of Psychometric Characteristics
Normative Sample
Factor Variance
Similar Factor Structure:
Probable fairness but bias
may be evident
Different Factor Structure:
Evaluate for instrument
bias and fairness

for demonstrating evidence of test content. In this respect, counselors who utilize assessment instruments should pay particular attention to item development. Issues of test bias and fairness may be inherent due to the theoretical framework from which an instrument was developed. For example, in the initial development of the Beck Depression Inventory in 1961, items were developed without any guiding theory of depression. “The 21 symptoms and attitudes chosen by Beck et al. (1961) for inclusion in the BDI were based on the verbal descriptions by patients and were not selected to reflect any particular theory of depression” (Beck Steer, & Brown, 1996, p. 2). However, through Beck’s involvement in cognitive behavioral therapy and refinement of the various editions of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM), the most recent iteration of items on the BDI-II is theoretically derived and related to the DSM-IV (Beck et al., 1996). With respect to test bias and fairness, items derived without theory and formulated based on subjective experiences expressed by patients in the early 1960s (a likely homogenous group), may have led to a widely used instrument that lacked generalizability across a variety of groups. As later versions (i.e., BDI-II) were developed with theory at the forefront of item development, a more generalizable instrument likely was generated.

Counselors should take time to evaluate the item content and generalizability of an instrument. Within the methods section of a manuscript under the description of measures used in a study, or within the theoretical explanation of an instrument in a test manual, counselors should be able to ascertain the extent to which theory was used to develop items and content experts reviewed the items and their association with the utilized theory. For example, Hambleton (1984) developed the index of item-objective congruence to provide a method of item evaluation with respect to specific goals/constructs items were designed to measure. Content experts (i.e., reviewers) rate the extent to which items measure an identified goal or construct using the following scale: -1 for an item that clearly *does not measure* an identified goal or construct, 0 for an item that *somewhat measures* an identified goal or construct, or +1 for an item that *clearly measures* an identified goal or construct. From these ratings a calculation can be performed to address the extent to which content experts agree an item is measuring an intended goal or construct. The index of item-objective congruence is a well-established method for evaluating evidence of test content and used in counseling literature (e.g., Balkin & Roland, 2007). However, methods for establishing evidence of test content by obtaining experts’ reviews of the items are also common. After developing items from a review of the literature, Kim, Soliz, Orellana, and Alamilla (2009) also surveyed members of a relevant professional organization and conducted focus group discussions. These procedures were outlined in Crocker and Algina (1986) on instrument development.

Evaluating the normative sample. The normative sample of a test, also referred to as a *norm group,* is the basis for score interpretation. Ideally, participants should come from a random sample, but true random sampling is quite rare in social science research given the need for volunteer participants and informed consent as well as assent from minors. Hence, identifying the extent to which scores of a norm group can be extended to individuals or groups should be based primarily on the representativeness of the sample. Issues of test bias and fairness may arise when scores from individuals or groups are compared to a normative sample that is qualitatively different or not representative.

For example, the BDI-II is likely one of the most popular instruments in measuring depression (Whiston, 2013). However, generalizability to non-White ethnicities may be limited. The BDI-II consisted of two samples: an outpatient sample (n = 500) that was 91% White and a college sample (n = 120) identified as "predominately White" (Beck, Steer, & Brown, 1996, p. 14) with no other demographic data presented. A further limitation may be the use of the BDI-II with adolescents. Beck et al. indicated an outpatient normative sample of 500 individuals ranging from 13 to 86 years of age with a mean age of 37.20 (SD = 15.91). The average age of participants ranges from 21.29 to 53.11. Adolescents (ages 13-17) likely comprised a small subset of the sample (i.e., less than 10% [n = 55] if age was normally distributed). An adolescent from an ethnic minority group will likely be compared to a small subset that has a high probability of lacking representativeness in terms of culture. While depression may indeed be a construct that is generalizable across many cultures and ethnic backgrounds, the extent to which symptoms are present is developmental as well. Therefore, when comparing scores comprised from a primarily adult population to adolescent scores on the BDI-II, generalizability may be limited.

When counselors evaluate the extent to which a test score is a fair, accurate representation for the client(s) or students, attention to normative data is pertinent. Counselors should take the time to familiarize themselves with the normative data on an instrument and make comparisons to individuals that completed the test under their administration. Such comparisons may provide evidence of test bias, which may or may not be an indicator of test fairness.

Evaluating factor invariance. Recall that testing for factor invariance involves the analysis of statistical tests, usually using latent variable modeling, to evaluate evidence that the properties and interpretations of test scores is similar across various groups (Dimitrov, 2011). When a different factor structure is evident from the scores between separate samples, an examination of the test content should be undertaken to consider whether the test is indeed biased and unfair. This was the case in the previous heuristic example using the Career Self-Efficacy Scale. Keep in mind that simply because a different factor structure exists does not mean the test is unfair, but the test may be measuring the construct differently or not at all.

Conversely, factor invariance may be substantiated when scores from separate samples yield a similar factor structure. Balkin et al. (2013) noted that the normative group for the Reynolds Adolescent Adjustment Screening Inventory (RAASI) consisted of a primarily White sample from two-parent homes with moderate to high incomes. When the factor structure of the RAASI was evaluated using Latino adjudicated youth from low socioeconomic status and single parent homes, a similar factor structure to that of the normative sample was identified. Therefore, while the normative sample may indeed be biased, the instrument is likely to be fair when used with the minority sample described.

Discussion

Sue, Arredondo, and McDavis (1992) strongly recommended "a multicultural approach to assessment, practice, training, and research" (p. 477); however, no meaningful model for evaluating multicultural factors in assessment is present in the literature. Our attempt at formulating a model for assessing the appropriateness for which assessment instruments may be utilized across cultures has meaningful implications to counselors, and especially counselors who work in school settings. The model for evaluating test bias and test fairness encourages counselors to be aware of theory involved in creating instruments and awareness of the basic psychometric qualities to validate the instrument. Counselors should be critical in their evaluation of the generalizability of the normative groups, as well as basic evidence of validity and reliability of scores. Moreover, counselors should be willing advocates for appropriate use of test scores, as well as identifying when cultural issues may pertain to the misinterpretation of test scores. Constructs commonly assessed in mental health disciplines, such as personality, behavior, and emotional states should be evaluated in conjunction with client culture (Whiston, 2013), and results of such tests should be interpreted cautiously when the normative sample is not representative of a client's culture.

Perhaps in no other area is the use of high-stakes testing more apparent than in the educational settings. As noted earlier, educational researchers often investigate differences in achievement test scores, despite the irrelevance of the construct to achievement. Professional school counselors (PSCs) are an obvious advocate for fair testing practices and procedures. There are many implications for school counselors to consider regarding the constructs that are irrelevant in assessment (Haladyna &

Downing, 2004; Helms, 2003) and how those irrelevant constructs are used for programming and placement practices such as pullouts, tutorials, accelerated instructional plans, or scheduling practices that reflect *course stacking* in the areas of reading and mathematics. Students who have their courses stacked may be placed in two reading or two math classes.

Some PSCs have major roles in their school's assessment and program coordination. These counselors should keep the results of the state assessment in perspective and train the faculty and staff on the constructs that are irrelevant in interpreting data, the non-cognitive variables (Sedlacek, 2004), and the complexities of multicultural and diversity issues (Moradi, Mohr, Worthington, & Fassinger, 2009) that could impact the scores of individual test takers. Consideration should be given to the "adverse testing conditions [that] may be a source of Construct Irrelevant Variance" (Haladyna & Downing, 2004, p. 21) and therefore, counselors should provide training that demonstrates how to replicate test preparation practices for all faculty and staff.

With regard to the amount of time provided by the schools for students to take state mandated assessments, PSCs should understand the potential ramifications of extending the amount of time for students to complete their tests (Haladyna & Downing, 2004) and to monitor their roles and involvement in the elimination of those students who are considered to be low performing from the overall testing population. This could lead to potential misrepresentation of a school's or district's actual achievement status (Haladyna & Downing, 2004). A final consideration may be that PSCs host parent education nights, with the purpose of discussing the importance of state assessment results. It should be explained that when in receipt of unfavorable score reports, parents should be made aware of the aforementioned variables that potentially impacted their children's success on state examinations.

PSCs should also be aware of state education agency practices and procedures for the administration of state-mandated assessments for students. PSCs should work with school administrators as well as state education agencies to provide policies that recommend non-biased used of assessment, and to define appropriate use for local school officials within K-12 settings. The implications of this suggest that state education agencies publish the norming information for which the test was developed in the test interpretation guides provided to school officials to use with their faculty, staff, parents, and community members. If the norming information is unavailable, these assessments should be deemed inappropriate.

Due to training in education and assessment, PSCs are ideal to provide relevant information to local school officials and district administrators on how these state assessments should be interpreted and utilized to make decisions about student programming. Limitations of assessment use and explicit statements depicting inappropriateness of use and practices to avoid should be provided. With respect to cultural, racial, physical ability status, and socioeconomic status, the validity of using a test to make decisions about a student from a status or background different from the test development sample may be challenged if the test appears to assess constructs related to background diversity (i.e., construct-irrelevant variance) rather than the construct defined as the stated purpose of the test (Helms, 2003).

With respect to all counseling professionals, we question the use of ethnicity as a comparative factor in addressing achievement test scores, because such a comparison undermines the principle of a valid test. If factor invariance is substantiated through the validation of a measure, then comparisons between ethnic groups are nonsensical.

References

American Counseling Association (2005). *ACA code of ethics*. Alexandria, VA: American Counseling Association.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Balkin, R. S., & Juhnke, G. A. (2014). *Theory and practice of assessment in counseling*. Columbus, OH: Pearson.

Balkin, R. S., Cavazos Jr., J. Hernandez, A. E., Garcia, R., Dominguez, D., & Valarezo, A. (2013). Assessing at-risk youth using the Reynolds Adolescent Adjustment Screening Inventory with a Latino/a population. *Journal of Addiction and Offender Counseling*, 30-39. doi: 10.1002/j.2161-1874.2013.00012.x

Balkin, R. S., & Roland, C. B. (2007). Re-conceptualizing stabilization for counseling adolescents in brief psychiatric hospitalization: A new model. *Journal of Counseling & Development, 85*, 64-72.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II manual*. San Antonio, TX: The Psychological Corporation.

Beutler, L. E., Brown, M. T., Crothers, L., Booker, K., & Seabrook, M. (1996). The dilemma of factitious demographic distinctions in psychological research. *Journal of Consulting and Clinical Psychology, 64,* 892-902. doi:10.1037/0022-006X.64.5.892

Council for Accreditation of Counseling and Related Educational Programs (2009). 2009 CACREP Standards. Alexandria, VA: Author.

Crocker, L., and Algina, J. (1986) *Introduction to classical and modern test theory*. Philadelphia, PA: Harcourt Brace Jovanovich College Publishers

Dimitrov, D. M. (2011). *Statistical methods for validation of assessment scale data in counseling and related fields*. Alexandria, VA: American Counseling Association.

Elwy, A., Ranganathan, G., & Eisen, S. V. (2008). Race-ethnicity and diagnosis as predictors of outpatient service use among treatment initiators. *Psychiatric Services, 59,* 1285-1291. doi:10.1176/appi.ps.59.11.1285

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues And Practice, 23,* 17-27. doi:10.1111/j.1745-3992.2004.tb00149.x

Hambleton, R. K. (1984). Validating the test scores. In R. K. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 199–230). Baltimore: Johns Hopkins University Press.

Helms J. (2003). *Fair and valid use of educational testing in grades K-12* [e-book]. Available from: ERIC, Ipswich, MA. Accessed November 19, 2012.

Kim, B. K., Soliz, A., Orellana, B., & Alamilla, S. G. (2009). Latino/a Values Scale: Development, reliability, and validity. *Measurement and Evaluation in Counseling and Development, 42,* 71-91. doi:10.1177/0748175609336861

Moradi, B., Mohr, J. J., Worthington, R. L., & Fassinger, R. E. (2009). Counseling psychology research on sexual (orientation) minority issues: Conceptual and methodological challenges and opportunities. *Journal of Counseling Psychology, 56,* 5-22. doi:10.1037/a0014572

Morley, C. P. (2010). The effects of patient characteristics on ADHD diagnosis and treatment: A factorial study of family physicians. *Morley BMC Family Practice, 11,* 1-10. doi: 10.1186/1471-2296-11-11

Nota, L., Ferrari, L., Solberg, V., Soresi, S. (2007). Career search self-efficacy, family support, and career indecision with Italian youth. *Journal of Career Assessment, 15,* 181-193. doi:10.1177/1069072706298019

Rabiner, D. L., Murray, D. W., Schmid, L., & Malone, P. S. (2004). An exploration of the relationship between ethnicity, attention problems, and academic achievement. *School Psychology Review, 33,* 498-509.

Sedlacek, W. E. (2004). *Beyond the big test: Noncognitive assessment in higher education.* San Francisco, CA: Jossey-Bass.

Solberg, V., Good, G., Nord, D., Holm, C., Hohner, R., Zima, N., Heffernan, M., & Malen, A. (1994). Assessing career search expectations: Development and validation of the career search efficacy scale. *Journal of Career Assessment, 2,* 111-123. doi:10.1177/106907279400200202

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York: Routledge.

Sue, D. W., Arredondo, P., & McDavis, R. J. (1992). Multicultural counseling competencies and standards: A call to the profession. *Journal of Multicultural Counseling and Development, 20,* 64-88. doi:10.1002/j.2161-1912.1992.tb00563.x

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, D. C.: American Psychological Association.

Whaley, A. (2004). Ethnicity/race, paranoia and hospitalization for mental health problems among men. *American Journal of Public Health, 94,* 78-81. doi:10.2105/AJPH.94.1.78

Whiston, S. C. (2013). *Principles and applications of assessment in counseling* (4th ed.). Belmont, CA: Brooks/Cole.