A Framework for Mixing Methods in Quantitative Measurement Development, Validation, and Revision : A Case Study

Russell Luyt

Abstract

A framework for quantitative measurement development, validation, and revision that incorporates both qualitative and quantitative methods is introduced. It extends and adapts Adcock and Collier's work, and thus, facilitates understanding of quantitative measurement development, validation, and revision as an integrated and cyclical set of procedures best achieved through mixed methods research. It also offers a systematic guide concerning how these procedures may be undertaken through detailing key "stages," "levels," and practical "tasks." A case study illustrates how qualitative and quantitative methods may be mixed through the use of the proposed framework in the cross-cultural content- and construct-related validation and subsequent revision of a quantitative measure.The contribution of this article to mixed methods research literature is briefly discussed.

Keywords

measurement development, measurement validation, measurement revision, mixed methods

A great deal of literature concerns procedures in the development, validation, and revision of quantitative instruments (e.g., Coaley, 2010; Hogan, 2007; Rust & Golombok, 2008). These, and many other texts, provide detailed and accessible coverage of such issues. This testifies to the maturity of the field. Yet an integrated framework for quantitative measurement development, validation, and revision, which incorporates both qualitative and quantitative methods, appears lacking. Onwuegbuzie, Bustamante, and Nelson (2010, p. 59) provide one of the few existing examples of work undertaken in this area and claim to provide a ". . . meta-framework for instrument development/fidelity and construct validation."

This article outlines an alternative framework for mixing methods in quantitative measurement development, validation, and revision through extending and adapting the work of Adcock and Collier (2001). These authors provide a useful framework that specifically considers measurement validation procedures and, in doing so, seeks to develop a methodological tool that can be used in both qualitative and quantitative research. The strength of this framework is evident in its clear and uncomplicated exposition. The authors avoid unnecessary and confusing jargon and procedural entanglements.

Yet their focus remains on measurement validation. Although the authors hint at related issues, such as measurement development and revision, these are neither fully developed nor discussed in detail. The current article extends and adapts their work to include an explicit discussion concerning the relationship between measurement development, validation, and revision. It describes this as a cyclical process best undertaken through mixed methods research, emphasizing the complementarity of qualitative and quantitative methods, and working from within a constructionist paradigm. A case study illustrates how qualitative and quantitative methods may be mixed in the validation and subsequent revision or development of a quantitative measure. A mixed methods approach of this kind offers advantages. This includes the common observation ". . . that the use of quantitative and qualitative approaches in combination provides a better understanding of the research problems than either approach alone" (Creswell & Plano Clark, 2007, p. 5). In particular, this approach provides a way in which to obtain more thorough validity evidence in support of measurement inferences than either method by itself. This offers a firmer foundation on which decisions regarding possible measurement revision or development can be made.

The case study presented in this article highlights some of the complexities in measurement development, validation, and revision through considering the cross-cultural measurement equivalence of the Male Attitude Norms Inventory–II (MANI-II; Luyt, 2005) in South Africa (SA). Measurement equivalence may be defined as degree to which ". . . research instruments used to collect data for the defined variables bear the same meanings and psychometric properties" (Tran, 2009, p. 67) cross-culturally. This is essential to make meaningful comparisons across cultural groups (van de Vijver & Tanzer, 2004). Yet insufficient cross-cultural validity evidence exists in support of the MANI-II's use among Afrikaans, English, and Xhosa speakers in the country.

Although much is written about the relationship between qualitative and quantitative methods (e.g., Griffin & Phoenix, 1994; Hammersley, 1992; Tashakkori & Teddlie, 2003), and combining methods in research is not new, mixed methods research is still a developing methodology (Brannen, 2009) or movement (Johnson, Onwuegbuzie, & Turner, 2007). A full review of this area is not possible here. It is sufficient to note that the obstacles in combining qualitative and quantitative methods exist on two main levels. The first is philosophical and the second practical (Brannen, 1992). These obstacles should not be underestimated. They have limited the extent to which authors have systematically made use of both in the development, validation, and revision of quantitative instruments.

Mixed methods research is best defined as a distinct methodology as well as a method. Methodologically, mixed methods research stresses the degree to which philosophical assumptions guide the collection, analysis, and mixing of qualitative and quantitative data (Creswell & Plano Clark, 2007). The incompatibility thesis, which suggests qualitative and quantitative methods are irreconcilable as a result of their underlying philosophical paradigms, is rejected (Tashakkori & Teddlie, 1998). This does not, however, imply paradigmatic agnosticism (Brannen, 2009). Various paradigms are said to be available to researchers, including for example, constructivism, postpositivism, and pragmatism (Creswell & Plano Clark, 2007), or alternatively, constructivism, positivism, postpositivism, pragmatism, and transformative (Teddlie & Tashakkori, 2009). Of these, pragmatism appears to be ascendant in the field at present (Johnson. et al., 2007; D. L. Morgan, 2007). Yet a constructionist paradigm (Gergen, 1985, 2005) is argued to provide the most suitable philosophical foundation when examining cross-cultural measurement equivalence. This paradigm broadly asserts that understanding of realities is sociohistorically dependent and (re)produced through social processes. It is consistent with an "indigenous" (van de Vijver & Tanzer, 2004, p. 122) or "particularizing" (Adcock & Collier, 2001, p. 530) approach to measurement, which recognizes that basic conceptual differences may exist between cultural groups in their understanding of a phenomenon. This approach is described in greater detail below. Other paradigms may be suitable in examining cross-cultural measurement equivalence, but should, at a minimum, account for such cross-cultural differences.

As a method, mixed methods research suggests practical procedures in the collection, analysis, and mixing of qualitative and quantitative data. Creswell and Plano Clark (2007), for example, describe three ways in which methods may be mixed:

> merging or converging the datasets by actually bringing them together, connecting the 2 datasets by having one build on the other, or embedding one dataset within the other so that one type of data provides a supportive role for the other dataset. (p. 7)

This article argues that merging qualitative and quantitative data sets proves beneficial in the proposed framework. In particular, a convergent parallel design is considered most suitable. Qualitative and quantitative approaches are seen to provide different, but complementary, data. These data are concurrently collected and analyzed. Findings are then merged and interpreted so as ". . . to measure overlapping but also different facets of (the) phenomenon, yielding an enriched, elaborated understanding" (Greene, Caracelli, & Graham, 1989, p. 258).

An emphasis on the complementarity of qualitative and quantitative methods differs from the traditional importance placed on convergence or triangulation in measurement validation literature. The concept of triangulation has a long and formative history in mixed methods literature (Campbell & Fiske, 1959; Denzin, 1970; Webb, Campbell, Schwartz, & Sechrest, 1966). Social scientists borrowed the concept from land surveying. The term originally referred to a technique in which multiple viewpoints improved accuracy of measurement. Similarly researchers in social science most frequently make use of triangulation in order to develop a clearer understanding of a phenomenon through ". . . convergence, corroboration, and correspondence of results across the different method types" (Caracelli & Greene, 1993, p. 196). Any single method is assumed to contain strengths and weaknesses. Weaknesses are supposedly mitigated through the use of more than one (Jick, 1983; Ponterotto & Grieger, 1999). Their combined use increases confidence in findings when data are consistent. Inconsistent data, on the other hand, suggest that researchers explore possible sources of bias. Understanding of triangulation is firmly embedded within a broadly positivist paradigm that stresses the existence of an objective and universal social truth (Brannen, 1992; Neuman, 1997). Nagy Hesse-Biber (2010) asserts that triangulation adopts a positivistic perspective by default:

> This design is employed when a researcher seeks to validate quantitative statistical findings with qualitative data results. Yet the assumption underlying triangulation is the positivistic view that there is an objective reality in which a given truth can be validated. The most common type of sequential mixed methods design appears to place the qualitative study in a supportive role. (p. 14)

Yet this article's emphasis on the complementarity of qualitative and quantitative methods—seeking "enriched, elaborated understanding"—places equal value on both consistent and inconsistent findings (Brannen, 1992). This understanding is congruent with the constructionist. paradigm as well as an "indigenous" or "particularizing" approach to measurement. Mixing qualitative and quantitative methods is considered beneficial, not because they may increase our confidence in findings through consistency, but rather because they are able to capture multiple realities (Ponterotto & Grieger, 1999). New explanations, questions, and even hypotheses are able to emerge as a result (Wolff, Knodel, & Sittitrai, 1993). In this sense, neither method is superior to the other in describing a single reality. Rather both contribute toward the understanding, description, and exploration of multiple realities. It is, therefore, advantageous for researchers to be proficient in the use of qualitative, as well as quantitative, methods. Ponterotto and Grieger (1999) observe that

> The researcher who can "wear two hats," so to speak, shifting in sequenced and integrative fashion between small-group descriptive and large-group normative approaches, will be more effective and better able to capture the true complexity of the phenomenon under study. (p. 56)

An ability to shift between these approaches, through the aid of mixed methods, was particularly important in the case study described below. "Small-group descriptive" qualitative data provided a means through which to obtain content-related validity evidence concerning the appropriate cross-cultural use of a quantitative measure (i.e., evidence of its cross-cultural content equivalence). "Large-group normative" quantitative data, on the other hand, provided a means through which to obtain construct-related validity evidence of its appropriate cross-cultural use (i.e., evidence of its cross-cultural construct equivalence).

Introducing a Framework for Measurement Development, Validation, and Revision

Adcock and Collier's (2001) extended and adapted framework is presented in Figure 1. It depicts measurement development, validation, and revision as three interconnected stages. These contribute toward a cyclical process whereby instruments are developed and necessarily undergo (re)validation over time and context. Validation procedures determine whether validity evidence exists in support of an instrument's use, and, on this basis, whether it requires revision or whether a new instrument might better be developed. Each cycle involves four sequential levels. Upward and downward movement between these levels is made possible through a series of research tasks. The three interconnected stages are described below in terms of these levels and tasks.

Stage 1: Measurement development. Measurement development involves defining the background concept (i.e., Level 1), developing a systematic and explicit definition of it (i.e., Level 2), and devising indicators that operationalize this systematic definition (i.e., Level 3).

The background concept may be described as ". . . the broad constellations of meanings and understandings associated with a given concept" (Adcock & Collier, 2001, p. 531). This might alternatively be described as the superordinate concept in that it includes or subsumes a narrower or more specified set of subordinate concepts. This article argues that its understanding should be firmly embedded in broader theoretical debate. Measurement development is usually motivated by practical need, theoretical concerns, or the need to revise existing instruments. However, measures motivated by practical need alone often demonstrate theoretical problems (Hogan, 2007). Theoretical grounding ensures that scholars are aware of the basic assumptions underlying the development of an instrument. It limits the extent to which inconsistencies and contradictions occur in its use, its future revision, or development of similar potential instruments. For example, if we were interested in developing a gender-related measure, we would find it useful to determine whether the construct of interest might best be theoretically described as assessing "gender orientation" or "gender ideology." The first considers gender as an individual

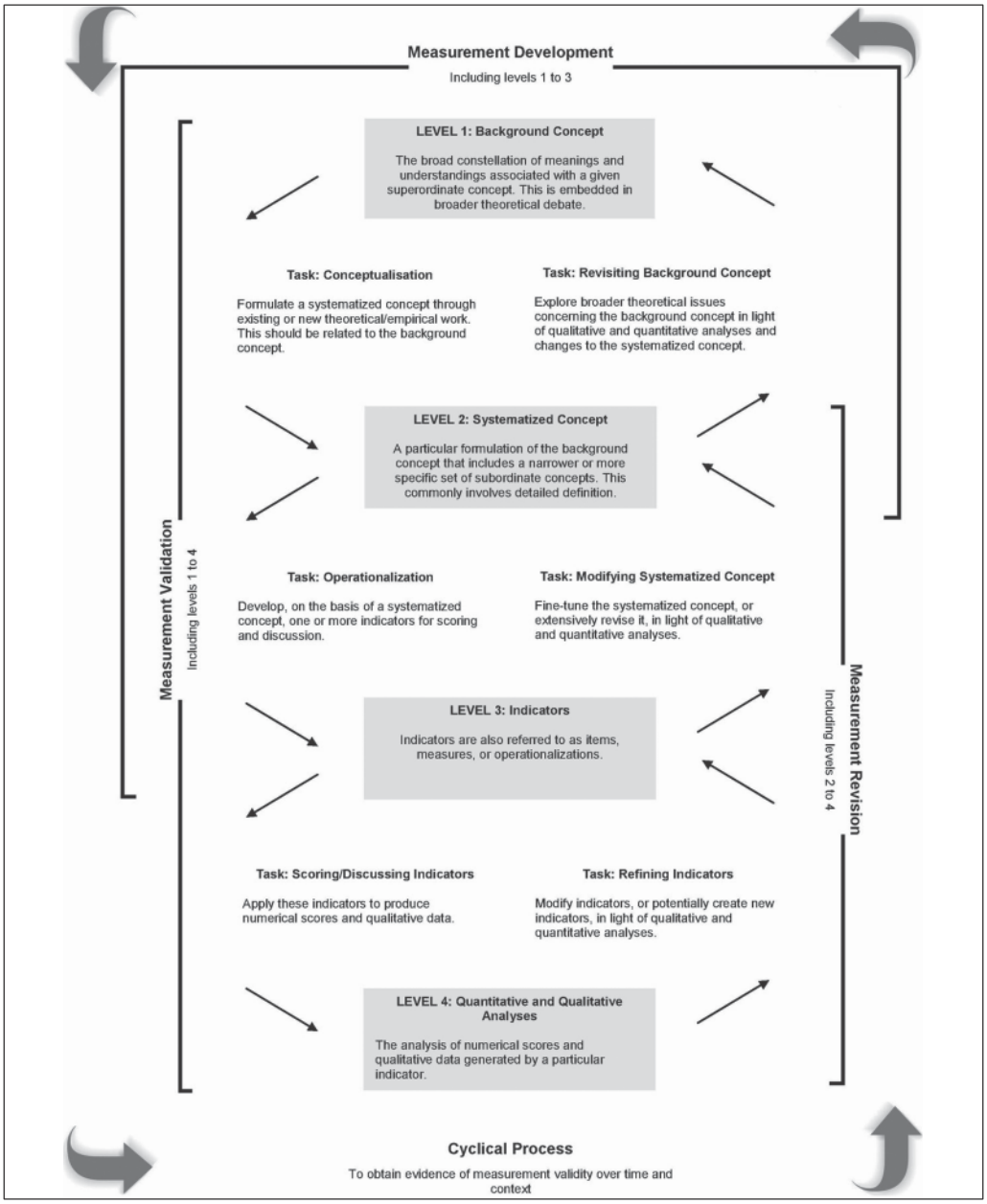

Figure 1. Measurement development, validation, and revision framework
Source: An extension and adaptation of Adcock and Collier (2001, p. 531).

property or trait and is frequently assessed through self-concept ratings. The second understands gender as a social norm and is commonly assessed through attitude scales (Thompson & Pleck, 1995, p. 130).

Adcock and Collier (2001) describe the systematized concept as a ". . . specific formulation of a concept used by a given scholar or group of scholars" (p. 531). However, this definition does not distinguish clearly enough between the background and systematized concept. The latter

might better be described as a specific formulation of a concept informed by existing or new theoretical/empirical work. It includes a number of narrower or more specified set of subordinate concepts that underlie the background concept. Detailed understanding of theory will suggest specific emphases in the development of the systematized concept. If, for example, the construct of interest is best explained as assessing "gender ideology," as described above, then literature indicates that this should be conceptualized as multidimensional among other things (Thompson & Pleck, 1995).

Existing or new empirical work might also aid in the formulation of the systematized concept and hence strengthen content-related validity evidence. There are various useful means through which this can be achieved. Thorough reading of relevant literature is essential. When doing so, scholars might first develop criteria and questions that guide their reading of material, and also consider developing a coding system that enables them to manage information in a systematic fashion. A careful record should be kept of this process as this serves as useful evidence of validity (G. A. Morgan, Gliner, & Harmon, 2001; Pett, Lackey, & Sullivan, 2003; Yun & Ulrich, 2002). Qualitative methods, such as focus groups, also aid in the formulation of the systematized concept. Focus groups have become an increasingly popular qualitative method in social scientific research (Bertrand, Brown, & Ward, 1992). Despite this, O'Brien (1993) notes that there is little literature concerning their worth in measurement development. A notable exception includes the work of Nassar-McMillan and Borders (2002) who discuss the role of focus groups in item development and refinement. The case study described below contributes toward this. A panel of judges may also usefully be appointed to review the adequacy of the systematized concept (G. A. Morgan et al., 2001; Yun & Ulrich, 2002).

The final task in measurement development is the operationalization of the systematized concept so as to produce indicators, items, or measures. Neuman (1997) defines this as ". . . a process of developing an operational definition for the construct" (p. 136). In practice, this requires developing indicators that capture the meaning of the systematized concept. These form the instrument and serve as the basis for scoring (G. A. Morgan et al., 2001). It is worthwhile to remember that instrument scores are only as good as the item from which they are derived (Coaley, 2010). Thankfully guidance concerning the development of indicators is abundantly available. This includes detailed advice concerning, for example, their wording, structure, and number; response format options; and so on (Pett et al., 2003). The instrument's theoretical inclination may again prove important at this stage. For example, measures of "gender ideology" are most appropriately argued to make use of third person statements that are absolute rather than relative (Levant et al., 1992; Thompson, Pleck, & Ferrera, 1992).

Stage 2: Measurement validation. Measurement validation involves assessment of whether the numerical scores and/or qualitative data (i.e., Level 4) obtained from indicators (i.e., Level 3) can meaningfully be interpreted in terms of the systematized (i.e., Level 2) and background concept (i.e., Level 1).

Traditionally measurement validity has been defined as the extent to which an instrument measures what it is intended to measure. Instruments are supposedly either valid or not valid. There are a number of procedures used for establishing this. These are commonly referred to as "types of validity." They include content, construct, and criterion validity. Each is seen to establish an independent and static property of an instrument (Messick, 1995; G. A. Morgan et al., 2001; Yun & Ulrich, 2002). Furthermore, any one is regarded as sufficient means with which to establish measurement validity (Adcock & Collier, 2001). However, this definition has been criticized over the past 20 years (Yun & Ulrich, 2002).

Critics argue that traditional understanding of measurement validity is flawed. Understanding validities as establishing independent and static properties of an instrument disregards the extent to which the appropriate interpretation of test scores is context specific. Measurement validit

has alternatively been defined as the degree to which theoretical and empirical evidence support the interpretation of test scores in a particular population and setting (Hogan & Agnello, 2004; G. A. Morgan et al., 2001; Yun & Ulrich, 2002). The difference between these definitions is made clear by Adcock and Collier (2001) who argue, ". . . the various procedures for assessing measurement validity must be seen, not as establishing multiple independent *types of validity*, but rather as providing different *types of evidence for validity*" (p. 530).

These authors do not assume that content, criterion, and construct validity established independent and static properties of an instrument. Rather they espouse what has been described as the unitary concept of validity. This suggests that only one type of validity exists. Thus, the use of a single "type of validity" is considered insufficient. Furthermore, the contextual specificity of measurement validation means that test score interpretation can only be generalized to settings and populations that are similar to the context in which evidence for validity was originally obtained. If researchers wish to use an instrument in a new context then they are obliged to provide evidence that the interpretation of test scores is appropriate. Measurement validation is therefore never complete (Messick, 1995; G. A. Morgan et al., 2001; Yun & Ulrich, 2002). It ". . . is an ongoing process because of the dynamic interaction among test participants, instrument, context, and purpose of the measurement" (Yun & Ulrich, 2002, p. 34). This underscores the cyclical nature of measurement development, validation, and revision.

A wide range of literature stresses the importance of reporting measurement validity and provides useful standards in this respect (e.g., American Educational Research Association [AERA], 2006; AERA, American Psychological Association, & National Council on Measurement in Education, 1999). However, research largely fails to provide necessary evidence for validity. This may be exacerbated by adherence to the traditional definition of measurement validity (Adcock & Collier, 2001). Detailed evidence for measurement validity is clearly important when adopting the unitary concept of validity. Collins, Onwuegbuzie, and Sutton (2006) describe a useful conceptual framework for areas of validity evidence. This framework subsumes traditional concepts of validity under the unitary concept: content-related (e.g., concurrent and predictive validity), construct-related (e.g., face, item, and sampling validity), and criterion-related validity (e.g., substantive, structural, convergent, and discriminant validity, etc.). These areas provide different types of validity evidence that support the interpretation of test scores in a particular context. The authors note that this framework is especially relevant when assessing validity evidence concerning the use of quantitative instruments.

Stage 3: Measurement revision. This involves determining whether indicators (i.e., Level 3) and the systematized concept (i.e., Level 2) require modification in light of the obtained numerical scores and qualitative data (i.e., Level 4). Adcock and Collier (2001) note that any revision of the systematized concept should not fundamentally alter it; it should rather only be extended. Furthermore, measurement revision should never involve the modification of the background concept. Broader disputes concerning this concept, and its associated theory, should only occur when it is deemed necessary to develop an entirely new instrument. A three-step procedure in measurement revision, which merges qualitative and quantitative findings, is outlined below.

Measurement Validation and Revision of the MANI-II: A Case Study

A case study illustrates how qualitative and quantitative methods may be mixed in the cross-cultural validation and subsequent revision of a quantitative measure. The procedure is guided by the framework for measurement development, validation, and revision described above.

The MANI-II may be described as a multidimensional measure of masculinity ideology. The inventory's background concept is theoretically based on the notion of gender as a social norm. This assumes that gender is defined through changing cultural understandings and associated

Table 1. The Systematized Concept of the Male Attitude Norm Inventory–II, Including its Five Dimensions and Their Underlying Concepts

	Dimensions				
	Sexuality	Toughness	Individualism	Status	Homophobia
Underlying concepts	Objectification of sex	Discomfort tolerance	Assertive activity	Achievement management	Homophobic ostracism
	Sexual control	Emotional detachment	Level-headed practice	Career management	Homophobic violence
	Sexual performance	Self-containment	Male independence	Resource management	Anti-homoerotic practice
	Masculine practice	Physical endurance	Interpersonal dominance	Power management	Homophobic avoidance

standards. It is perhaps unsurprising, therefore, that the inventory seeks to examine variable attitudes toward masculinity. Its systematized concept includes five theoretical dimensions that are thought to underlie understanding of traditional masculinity in SA. Each dimension is further defined through four key concepts (see Table 1). The systematized concept was informed through a review of germane empirical and theoretical literature as well as qualitative and quantitative analyses. Underlying concepts served as the basis for the development of 40 prescriptive statements. Participants are asked to indicate their agreement or disagreement with these statements along a 5-point Likert-type response format. Their responses are believed to index the extent to which they endorse traditional norms of masculinity in SA.

Available content- and construct-related validity evidence lends support for the use of the MANI-II (English Version) in SA (Luyt, 2005). However, the Afrikaans, English, and Xhosa version of the MANI-II are merely assumed to be equivalent. Insufficient cross-cultural content- and construct-related validity evidence exists to warrant such an assumption. Qualitative and quantitative analyses were undertaken to determine whether such evidence exists. As described above, a convergent parallel design was considered most suitable, where data sets were concurrently collected and analyzed, findings merged, and ultimately interpreted (QUAN + QUAL).

The findings of the case study indicated that different language versions of the MANI-II do not demonstrate sufficient cross-cultural measurement equivalence. Qualitative findings did not provide evidence in support of the measure's cross-cultural content equivalence. In particular, cross-cultural item bias was evident. van de Vijver and Tanzer (2004) note that this ". . . is most frequently caused by poor item translation, ambiguities in the original item, low familiarity/appropriateness of the item content in certain cultures, or influence of cultural specifics such as nuisance factors or connotations associated with item wording" (p. 127). Quantitative findings only provided partial evidence in support of the measure's cross-cultural construct equivalence. That is to say, ". . . the same construct is (not) measured across all cultural groups" (van de Vijver & Tanzer, 2004, p. 121).

Ben-Porath (1990) notes that analysis of data collected from different cultural groups may indicate basic conceptual differences and, therefore, the need to develop separate measures for each. This is consistent with a constructionist paradigm and has been described as an "indigenous"

or "particularizing" approach to assessment which emphasizes the importance of developing separate measures that are sensitive to specific contexts and populations. In light of findings, it was important to determine whether (a) the instrument should be revised, to enhance cross-cultural measurement equivalence or (b) to adopt a particularizing approach where new instruments should be developed for different cultural groups. Joint consideration of qualitative and quantitative analyses suggested that basic conceptual differences in the understanding of traditional masculinity do not exist between cultural groups. Thus, although development of new instruments was unnecessary (i.e., Levels 1-3), revision of the MANI-II was beneficial so as to improve its cross-cultural measurement equivalence (i.e., Levels 2-4).

Qualitative Analysis: Cross-Cultural Content-Related Validation of the MANI-II

Qualitative analysis was undertaken to determine whether sufficient cross-cultural content-related validity evidence exists to warrant the use of the MANI-II among different cultural groups in SA—hence the cross-cultural content equivalence of different language versions. The *Standards for Educational and Psychological Testing* suggest that this form of

> . . . validity evidence can be obtained from an analysis of the relationship between a test's content and the construct it is intended to measure. Test content refers to the themes, wording, and format of the items, tasks, or questions on a test. (AERA et al., 1999, p. 11)

If insufficient evidence of cross-cultural content-related validity was found to exist, it was hoped that the analysis would indicate whether the instrument should be revised, or new instruments should be developed for different cultural groups.

Qualitative analysis relied on data from six focus groups. Each represented fairly homogeneous subpopulations of SA society and were grouped on the basis of spoken home-language (i.e., Afrikaans, English, and Xhosa), self-defined race (i.e., Black African, Colored, and White), and attained education level (i.e., primary, secondary, and tertiary). Age was not included as a formal stratifying criterion. Nevertheless an attempt was made to involve men of different ages in groups.

The composite stratifying criterion of spoken home-language and self-defined race is believed to offer a proxy indicator of ethnicity. Formal education level similarly served as a suitable substitute indicator of socioeconomic status. Unlike quantitative methods, which attempt to achieve representativeness through appropriate sampling, qualitative studies rely on theoretically led sampling. As such, data do not result in generalizable findings but rather offer rich insight into areas of research concern. This contributes toward theory and hypothesis building (Kelle & Laurie, 1995). These issues were of central concern in this analysis.

Discussion was designed to assess whether the items held the same meaning for people within and between groups. It was hoped that this focus would render results which would indicate whether or not the items adequately operationalized the systematized concept they were developed to assess. If items were not found to adequately operationalize this concept then an attempt was made to determine whether this was because of their weak wording and/or translation or insufficient shared conceptual understanding. This procedure is suggested by Nassar-McMillan and Borders (2002, ¶ 6) who note that although there are no clearly defined rules for the use of focus groups in measurement development, they aid in comparing ". . . groups' reactions to the same concepts."

Conversation was transcribed in full and, where necessary, back-translated into English (Brislin, 2000). An iterative inductive–deductive thematic analysis was then undertaken (Braun & Clarke, 2006). Literature concerning conceptual and methodological concerns in cross-cultural assessment (Tran, 2009; van de Vijver & Poortinga, 2005; van de Vijver & Tanzer, 2004) as well

as the measurement of masculinity ideology (Thompson et al., 1992) contributed toward informing this. Three relevant higher order themes emerged. The theme of *conceptual evidence* included all text references in which within- or between-group disagreement surfaced concerning the general idea or notion captured by either the measure or its items. Conceptual evidence of content-related validity is therefore situated between Levels 1 and 2 (i.e., background concept → conceptualization → systematized concept) as well as between Levels 2 and 3 (i.e., systematized concept → operationalization → indicators).

What did analysis of conceptual evidence between, for instance, Levels 1 and 2 entail? This involved determining whether the systematized concept had been adequately and appropriately formulated through reasoning concerning the background concept. The MANI-II's background concept has been defined as the social norms of traditional masculinity. The systematized concept builds on this definition. In doing so it includes five theoretical dimensions of traditional masculinity (i.e., "Sexuality," "Toughness," "Individualism," "Status," and "Homophobia") as well as their underlying concepts (e.g., "self-containment," "physical endurance," and "homophobic violence"). Data indicated that, for example, the conceptualization of the systematized concept might overemphasize homosexuality.

Often, when asked to provide general feedback concerning the measure, participants commented it seemed to be assessing attitudes toward homosexuality:

Extract 1 (English-White men above grade 9 education)

P1: Quite clearly it's about homosexuality. And in my view that's a popular subject at the moment, I'm getting a bit bored with it, really . . . (inaudible). . . there's nothing you can do about it. . .
Int: But what would you first think, looking at this questionnaire, that you would think, you would assume the objective to be?
P3: I, from this I would, er, I understood the thrust of the questions to assess the, um, view of males, participating, um, of homosexuality. Male homosexuality. Lying in the same bed, and things like that, taking different . . .
Int: Would you all agree with that? Was that the impression that you got?
P1: Ja.

The claim that the measure appears to be assessing attitudes toward homosexuality is surprising given that only 8 out of a total of 40 items do so. Additionally Item 6 ("Being called a 'faggot' is one of the worst insults to a man") was the first "homophobic" concept to appear and Item 37 ("A father should be embarrassed if his son his gay") the last. It is therefore unlikely that item order served to markedly bias participant understanding of the inventory.

Qualitative research exploring masculine representation in SA (Luyt, 2003, in press) suggests that (hetero)sexuality is central in the construction of contemporary masculinity. Masculinity ideology literature supports this view. Both Levant et al. (1992) and Thompson and Pleck (1995) argue crucial dimensions, such as homophobia and sexual attitudes, are often omitted or underrepresented in related measures. It was for this reason that the additional dimension of "Sexuality" was incorporated in the MANI-II so as to stress the importance of performative sexuality (Luyt, 2005).

Yet it is interesting to note that participants did not suggest that the inventory appeared biased toward issues of sexuality in general (i.e., the dimensions of "Sexuality" and "Homophobia" taken together). This would have been easier to account for given that 40% of the questionnaire includes sexuality-related content (i.e., 16/40 items). Rather participants understood the inventory as specifically exploring attitudes toward homosexuality. This response may reflect homophobic defense on the part of participants who stressed that they didn't "have much to do with gay people" (Line 2):

Extract 2 (Afrikaans-Colored men equal and below grade 9 education)

```
01  P3:  Most of the questions are difficult, because we
02       don't have much to do with gay people.
03  Int: Okay.
04  P3:  To me, especially . . .
```

The second theme of *contextual evidence* included all text references in which between-group or cultural disagreement surfaced concerning the general idea or notion captured by items. Contextual and conceptual evidence are therefore similar. However, the current theme differs in its added emphasis on the contextual specificity of items. This specificity only becomes apparent through between-group comparison. It is argued that contextual evidence of content-related validity is always situated between Levels 2 and 3 (i.e., systematized concept → operationalization → indicators).

What did analysis of contextual evidence between these levels involve? As noted, this required determining whether the indicators had been adequately and appropriately developed on the basis of the systematized concept. Item 11 ("Men should be able to sleep close together in the same bed") provides an example. Although all participants seemed to endorse the concept of traditional masculine "anti-homoeroticism," they nevertheless claimed that the item lacked sufficient contextual specificity, where in some cases "sleep(ing) close together in the same bed" would be quite permissible. This is made clear by English men who maintained "it depends on the situation" (Lines 5-6):

Extract 3 (English-White men above grade 9 education)

```
01   P2:  I've got no opinion on it, because I mean you
02        could be suffering, in the Antarctic, and you
03        might have to sleep together very closely to keep
04        warm, to survive.
05   P3:  I had no opinion, because it depends on the
06        situation, and I suppose if folk want to sleep,
07        you know I personally don't like to sleep close
08        to anyone that's in the bed, male or female, I
09        like to have my . . . (stretches arms out).
```

In this extract it is suggested that men may have to "sleep together very closely to keep warm, to survive" (Lines 3-4), in the "Antarctic" (Line 2). Examples, relating to situations in which male physical closeness would be acceptable, differed according to participant socioeconomic status. Xhosa males saw sleeping "close together in the same bed" with another man as legitimate as long as they "face opposite directions" (Lines 11-12). This reflects economic reality where these men often only "have one bed" (Line 15):

Extract 4 (Xhosa-Black African men equal and below Grade 9 education)

P4: (Inaudible) . . . like if may be he is your younger brother and there is something, you see? They will be busy, I will sleep with him and no one is going to have a problem . . .
P6: I would understand it when they say they were close to each while fast asleep. There is a difference when we are close to each other in back-to-back if the bed is small.
P3: Listen, you see I understand the way you put it only if you are my brother. If you are not my brother you can sleep in the feet, and we will face opposite directions . . . (Participant 3 pointing at opposite directions) . . . in the feet, you see . . .
P6: (Talking) . . . if for example sleeping there is your family, and you have one bed and your father is in rural areas, you see? You are with your friend, now he does not feel like going to Gugs . . . (refers to the local township of Gugulethu). He says, no I will sleep over here. Now are you trying to tell me that he is going to sleep in the bed and you sleep down on the floor? And what is the person on the other bed going to say when s/he sees you sleeping down?
P7: I will sleep with him, but I will take my blanket and give him his.

Clearly, this extract suggests that Item 11's content should be altered so as to reduce its contextual ambiguity. This needs to be achieved without resorting to emotive language that may bias participant response. Accordingly revision would involve a minor change to item content (e.g., "Men should be able to sleep *intimately* together in the same bed"). A simple transformation, such as this, accomplishes greater specificity without developing an entirely new indicator.

Last, the third theme of *semantic evidence* included all text references in which disagreement emerged concerning the meaning of language used by specific indicators. It is argued that semantic evidence of content-related validity is always situated between Levels 2 and 3 (i.e., systematized concept → operationalization → indicators).

What did analysis of semantic evidence between these levels entail? Once again this required determining whether the indicators had been adequately and appropriately developed on the basis of the systematized concept. Word use in seven items appeared questionable. For example, Afrikaans individuals agreed that the use of "admirable" ("loofwaardig"; Line 1) in Item 23 ("It is admirable for a man to take the lead when something needs to be done") was inappropriate:

Extract 5 (Afrikaans-Colored men above grade 9 education)

P2: . . . 'admirable' ('loofwaardig'). I don't understand the word that well. I understand the question, but not . . .

```
04   Int:  All right . . . the word to be specific.
05   P2:   Yes . . .
06   Int:  As it's here . . . in the sentence, and the way in
07         which you understand it. What do you think it tells
08         you? Just in your own thoughts.
09   P2:   He must be heroic . . .
10   Int:  Okay. Any comments?
11   P1:   I agree with him. To me it means the man should . . .
12         he should take the first step in leading. Almost
13         like an example.
```

The words "admirable" and "heroic" (Line 9) convey different meaning in English. It is unsurprising therefore that the participants find the statement, "It is heroic for a man to take the lead when something needs to be done," peculiar. This provides a good example of poor translation. Here the translation of "admirable" led to the unsuitable use of "loofwaardig," which participants struggled to meaningfully interpret within the item.

In sum, qualitative findings did not provide evidence in support of the measure's cross-cultural content equivalence. That is to say, although basic conceptual differences did not emerge between cultural groups as evidenced in the background (i.e., Level 1) or systematized concept (i.e., Level 2), cross-cultural item bias was evident. This suggested that, although it was unnecessary to develop separate measures for each cultural group, revision needed to take place so as to improve the measure's cross-cultural content equivalence. Thus, refinement of the indicators (i.e., Level 3) and modification of the systematized concept (i.e., Level 2) was required (Table 2). A summary of these findings across language groups and their relationship to the proposed framework for measurement development, validation, and revision appears in Figure 2.

Table 2. The Revised Systematized Concept Including its Six Dimensions and Their Underlying Concepts

	Dimensions					
	Sexuality	Toughness	Independence	Status	Responsibility	Homophobia
Underlying concepts	Sexual objectification	Emotional containment	Assertive behavior	Behavior management	Duty	Homophobic ostracism
	Sexual control	Emotional denial	Achievement management	Career management	Dependability	Homophobic violence
	Sexual self-appraisal	Self-containment	Self-actualization	Resource management	Self-sacrifice	Anti-homoerotic behavior
	Sexual other-appraisal	Physical tenacity	Interpersonal dominance	Power management	Accountability	Homophobic avoidance

Quantitative Analysis: Cross-Cultural Construct-Related Validation of the MANI-II

Quantitative analysis was undertaken to determine whether sufficient cross-cultural construct-related validity evidence exists to warrant the use of the MANI-II among different language groups in SA—hence the cross-cultural construct equivalence of different language versions. The *Standards for Educational and Psychological Testing* suggest that " . . . (a)nalyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are

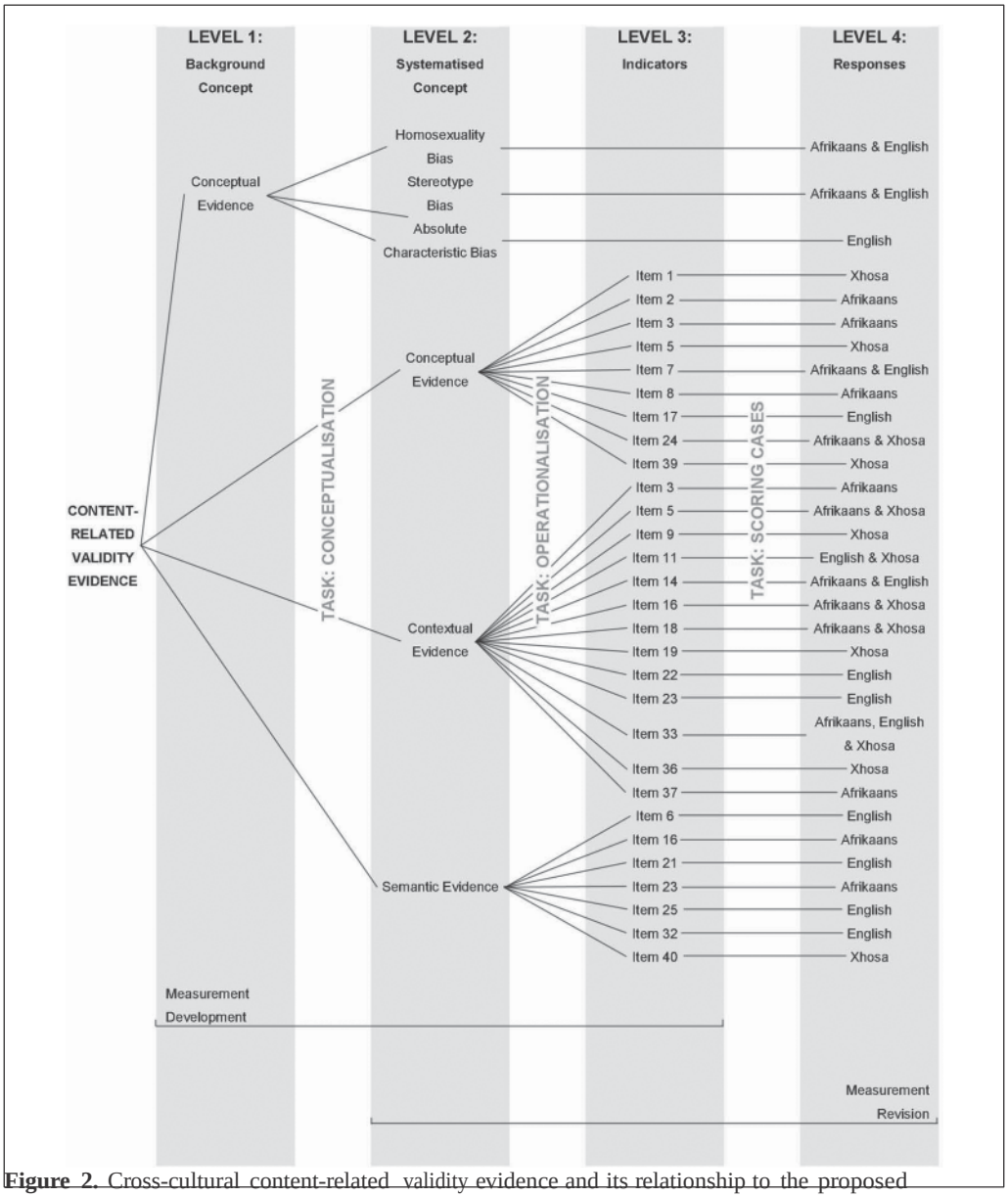

Figure 2. Cross-cultural content-related validity evidence and its relationship to the proposed framework for measurement development, validation, and revision

based" (AERA et al., 1999, p. 13). If insufficient evidence of cross-cultural construct-related validity was found to exist, it was hoped that the analysis would indicate whether the instrument should be revised, or entirely new instruments should be developed for different cultural groups.

Quantitative analyses relied on data from two independent samples as a means of data triangulation (Denzin, 1978). Sample A comprised male students (n = 639) from SA. Questionnaires were completed in Afrikaans (n = 248), English (n = 228), and Xhosa (n = 176). Sample B included a selection of SA residents (n = 1,597) stratified by age, spoken

home-language (i.e., Afrikaans, English, and Xhosa), and self-defined race (i.e., Black African, Colored, and White). Questionnaires were completed in Afrikaans (n = 434), English (n = 890), and Xhosa (n = 273).

Quantitative analyses sought to assess the extent to which the instrument's underlying construct carries the same meaning within different cultures. Evidence used to support an instrument's construct equivalence originates through cross-cultural comparison of either its internal structure or its relationship to meaningful external correlates. Both are important (Allen & Walsh, 2000). Current analyses specifically explored the equivalence of the MANI-II's internal structure across Afrikaans, English, and Xhosa speakers.

A number of methods exist for assessing construct equivalence. These include cluster analysis, multidimensional scaling, and structural equation modeling. However, factor analysis remains the preferred method for determining the equivalence of an instrument's internal structure across different cultural groups. In this case, construct equivalence was operationally defined as factorial invariance (Allen & Walsh, 2000; Ben-Porath, 1990; Floyd & Widaman, 1995). A particular method of determining factorial invariance, known as replicatory factor analysis, was applied. This includes a number of key steps: First, the instrument should be appropriately translated and then administered to participants in the new culture. Second, data should be factor analyzed. The same procedures for estimating communalities and rotation, which were used in the analysis of the original data, need to be applied. Third, the number of factors extracted should be restricted to the number having emerged from the original data. Coefficients of congruence may be calculated. These provide a formal test of factor similarity. These successive steps allow a comparison (Table 3) to be made between the factor structures that emerged from Luyt's (2005) original study based on the English version and later analyses of Afrikaans, English, and Xhosa versions (Ben-Porath, 1990).

Table 3. Factorial Invariance Between, and Percentage Variance Explained by, the Target Versus the Replicatory Factors

	Target Factors			Total
	Factor 1 "Toughness"	Factor 2 "Success"	Factor 3 "Control"	
Original study				
English	% var. = 11.01	% var. = 10.61	% var. = 9.82	% var. = 31.44
Sample A				
Afrikaans	**f** = .95, % var. = 10.08	**f** = .91, % var. = 7.85	**f** = .93, % var. = 10.28	% var. = 28.21
English	**f** = .89, % var. = 8.78	**f** = .96, % var. = 9.75	**f** = .93, % var. = 13.78	% var. = 32.29
Xhosa	**f** = .89, % var. = 10.50	**f** = .91, % var. = 12.46	**f** = –.53, % var. = 5.33	% var. = 28.29
Sample B				
Afrikaans	**f** = .91, % var. = 14.54	**f** = .858, % var. = 8.98	**f** = .82, % var. = 5.06	% var. = 28.58
English	**f** = .96, % var. = 10.15	**f** = .91, % var. = 9.97	**f** = .96, % var. = 6.76	% var. = 26.88
Xhosa		**f** = .82, % var. = 11.00		

Note: **f** = Tucker's phi; % var. = percentage variance.

Findings, as assessed by Tucker's coefficient of congruence (Tucker's phi), suggested that Xhosa participants did not appear to understand the instrument's underlying construct in the same way as Afrikaans and English participants. This formal test of factor congruence provided strong to moderate support for factorial invariance among Afrikaans and English samples but moderate to weak support in the case of the Xhosa samples. Findings were difficult to interpret at times. This was, particularly, so with regard to Xhosa analyses where, for example, only a

single factor was interpretable in the case of Sample B. In most instances, total item variance accounted for by replicatory factor solutions appeared in close approximation to that explained by the original study.

In sum, quantitative findings only provided partial evidence in support of the measure's cross-cultural construct equivalence. Yet the reason for these mixed findings was equivocal. This may have indicated "idiosyncrasies of each culture" (van de Vijver & Tanzer, 2004, p. 122) where traditional masculinity is understood somewhat differently across cultural groups as applicable to the background (i.e., Level 1) and/or systematized concept (i.e., Level 2). It suggests that a particularizing approach be adopted where entirely new instruments are developed for different cultural groups. Yet when the results of the qualitative analysis were taken into consideration, it seemed unlikely that reasons for measurement inequivalence lay at a basic conceptual level, but rather that this was because of contextual or semantic issues. Cross-cultural item bias could have affected quantitative findings; this was evident, for example, in their sometimes poor interpretability. Therefore, revision of the MANI-II, through reconsideration of the systematized concept (i.e., Level 2) and indicators (i.e., Level 3), was deemed most appropriate so as to enhance cross-cultural construct equivalence. A summary of these findings across samples as well as language groups and their relationship to the proposed framework for measurement development, validation, and revision appears in Figure 3.

Merging Qualitative and Quantitative Analyses in Measurement Revision

Taken together, qualitative and quantitative analyses suggested that revision of the MANI-II would be beneficial so as to enhance cross-cultural validity evidence in support of its use across all three cultural groups in SA, and as such improve its cross-cultural measurement equivalence. The proposed framework for measurement development, validation, and revision revealed that this needed to include two tasks. These were situated between Level 2 and Level 4. First, indicators needed to be refined, or new indicators created, on the basis of qualitative or quantitative analyses. Second, the systematized concept needed to be modified on the basis of the refined or new indicators.

To merge qualitative and quantitative findings in the revision of the inventory a three-step procedure was adopted. This is outlined in Figure 4. There are no set guidelines concerning how to merge findings in a procedure of this kind. It is more usual for qualitative data to supplement key findings provided by quantitative data (Jick, 1983). Quantitative data were considered before qualitative data in the procedure outlined below. However, qualitative data were not considered merely supplementary. In this case, they provided unique (i.e., Step 3) as well as supplementary information (i.e., Steps 1 and 2) concerning possible item/instrument revision.

Task: Refining indicators. A summary table of the qualitative and quantitative evidence concerning the MANI-II's measurement validity was produced. This was helpful in providing an overview of findings. It facilitated the three-step procedure described above. An example of each step is useful.

Step 1 was relevant when evaluating Item 16 (i.e., "Men should feel embarrassed if they are unable to get an erection during sex"). That is to say, factor analytic evidence suggested that revision might be necessary. The item had an insubstantive factor loading across Samples A and B. Qualitative evidence provided a possible reason for this result. Afrikaans and Xhosa focus group participants argued that men should only feel embarrassed in certain situations, for example, when they were attempting to have intercourse with a new sexual partner. The item was revised accordingly (i.e., "Men should feel embarrassed if they are unable to get an erection with a new sexual partner").

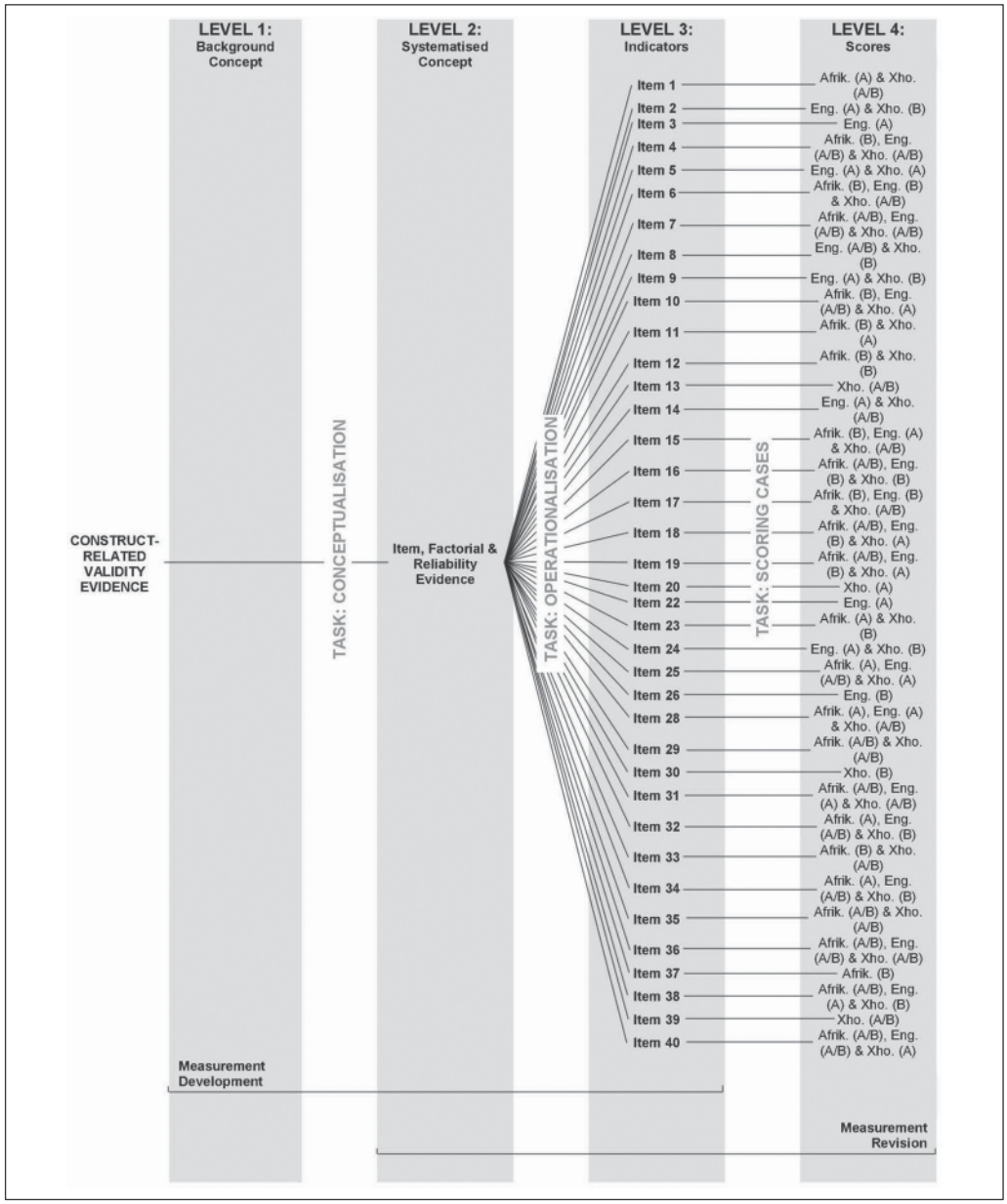

Figure 3. Cross-cultural construct-related validity evidence and its relationship to the proposed framework for measurement development, validation, and revision

Step 2 was applicable when considering Item 7 (i.e., “Men should think logically about problems”). Item analytic evidence suggested that revision might be necessary. Specifically, the item displayed an extreme mean and low standard deviation. Qualitative evidence revealed a possible reason for this result. Afrikaans and English participants maintained that the concept of logic does not typify masculinity but is rather a general societal expectation. The item was

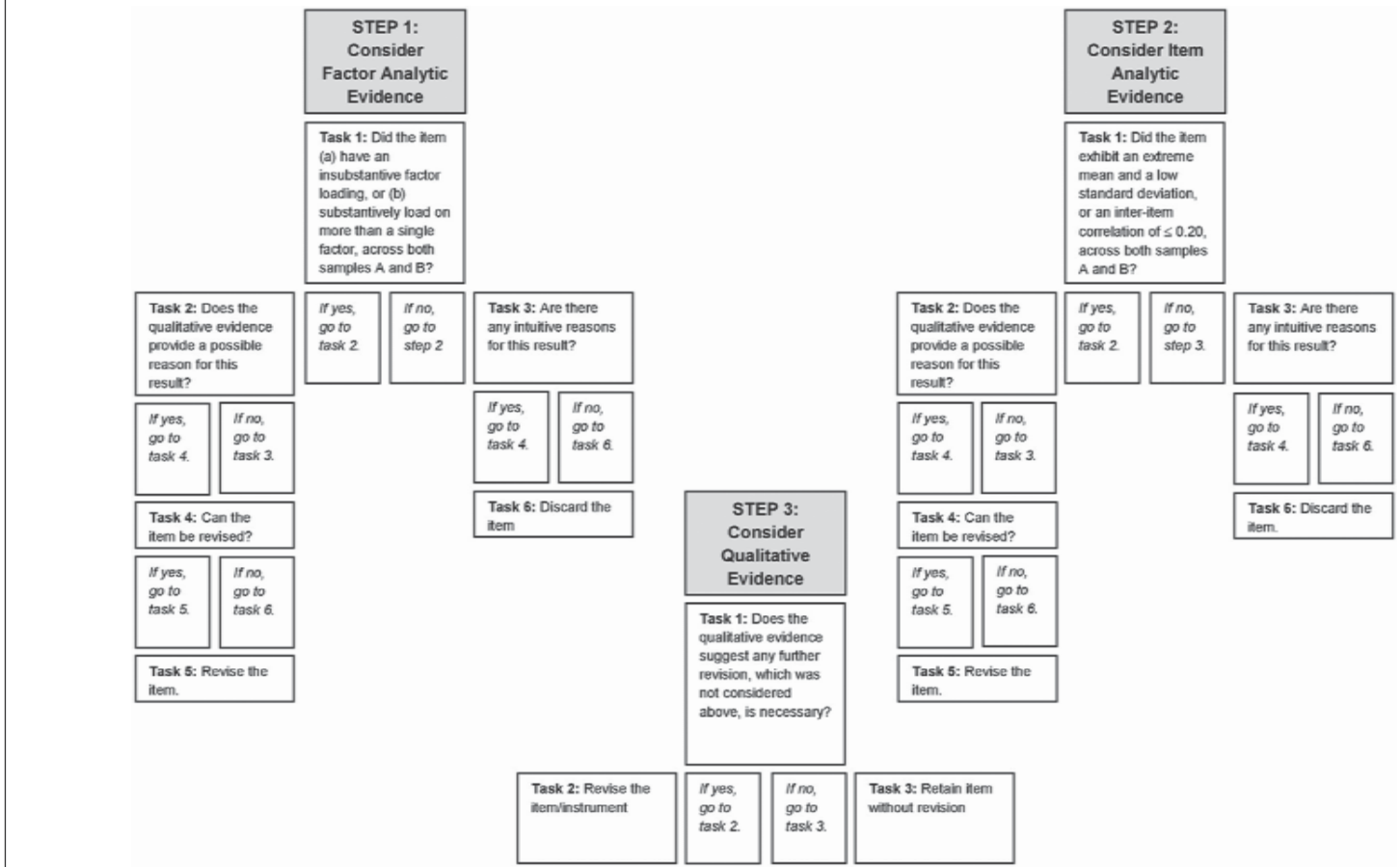

Figure 4. Three-step procedure adopted in instrument revision

discarded on the basis that it did not constitute an absolute gender characteristic as suggested by theory (Levant et al., 1992).

Step 3 was most appropriate in the case of Item 5 (i.e., "To be a man you need to be tough"). Qualitative evidence suggested that revision might be necessary. Afrikaans and Xhosa participants indicated that the item may be improved through emphasizing the difference between emotional as opposed to physical toughness. The item was revised. It now specifically operationalizes the concept of physical toughness (i.e., "To be a man you need to be physically tough").

In some instances, neither qualitative nor quantitative evidence emerged. This suggested that an item did not need to be refined (e.g., Item 30: "Gay men are not suited to many jobs"). These items were retained without revision. On other occasions evidence emerged suggesting that semantic revision was necessary. Translation into Afrikaans and Xhosa was required in cases where this resulted in the refinement of an English item. For example, in Item 25 (i.e., "A man should not worry about the future") focus group participants argued that the word "worry" is not characteristic of "real" men. The item was revised (i.e., "A man need not plan well in advance for the future"). Its translation into Afrikaans and Xhosa was therefore necessary. However, when semantic evidence emerged that resulted in the revision of Afrikaans or Xhosa items, this was limited to the language version in question. For example, Afrikaans focus group participants argued that the use of the word "verleë" in Item 16 (i.e., "Men should feel embarrassed if they are unable to get an erection during sex") was an inappropriate translation. That is to say, they argued that the word referred to feeling "ashamed" rather than "embarrassed." Translators also drew on colloquial language that flowed from group debate to produce more meaningful items.

Finally, qualitative evidence suggested that the inventory relies heavily on stereotypical definitions of masculinity. As has been noted in literature, this places limits on the extent to which agentic self-concept may be explored (Beere, 1990). However, this criticism was not considered relevant in the case of ideological assessment, which seeks to determine the extent to which individuals endorse traditional and hence often stereotypical notions of masculinity. The criticism that measures, such as the MANI-II, negatively reinforce an association between stereotypical and "real" notions of masculinity through stereotypical definition, is more worrying (Hoffman, 2001). It is argued that both traditional and alternative statements should be incorporated in measures of gender ideology. In this way, researchers can be more certain that participant rejection of traditional norms reflects an endorsement of alternative masculinity. Furthermore, a more even allotment of these statements would contribute toward a politics of change. That is to say, alternative notions of masculinity might also be reinforced. The MANI-II included 10 reverse-scored items. These constituted 25% of total item content. The revision process resulted in 17 reverse-scored items. This increased the number of alternative statements to approximately 35% of total item content.

Task: Modifying the systematized concept. The next task in measurement revision involved the modification of the systematized concept. The insights gained during item revision guided this task. Table 2 presents the outcome. The changes are obvious when the modified systematized concept is compared against the original systematized concept in Table 1. A number of examples are provided below that illustrate this process.

Two changes are evident at a dimensional level. First, the concepts underlying the dimension of "Individualism" in the original systematized concept (i.e., "assertive activity," "level-headed practice," "male independence," and "interpersonal dominance") were extensively modified (i.e., "assertive behavior," "achievement management," "self-actualization," and "interpersonal dominance"). These underlying concepts were considered indicative of independence rather than individualism. The dimension was therefore renamed. Second, qualitative findings (Luyt, 2003) highlighted the importance of responsibility in the definition of traditional masculinity. A new dimension of "Responsibility" was therefore included in the modified systematized concept.

This increased the number of dimensions from five to six. This represented a case of construct underrepresentation where the ". . . test fail(ed) to capture important aspects of the (background) construct" (AERA et al., 1999, p. 10).

In numerous cases, the concept operationalized by an item required modification once the item had been revised. For example, Item 16 (i.e., "Men should feel embarrassed if they are unable to get an erection during sex") and Item 18 (i.e., "It is not important for men to achieve orgasm during sex") operationalized the concept of "sexual performance." These items were revised. Item 20 (i.e., "Men should feel embarrassed if they are unable to get an erection with a new sexual partner") and Item 40 (i.e., "Men should not feel embarrassed if they ejaculate before being able to make love") are now considered to operationalize the concept of "sexual self- appraisal." Furthermore, item revision occasionally resulted in it operationalizing an existing but different concept. For example, Item 4 (i.e., "Men should share their worries with other people") originally operationalized the concept of "self-containment." It was revised and Item 48 (i.e., "A man should tell others when he is feeling depressed") is now considered to operationalize the concept of "emotional denial."

Finally, neither qualitative nor quantitative findings indicated the need to revisit the background concept. Findings of this kind would have suggested fundamental theoretical tensions. In such cases, theoretical reconsideration would be necessary, where qualitative findings would likely contribute most. This would ultimately guide the development of an entirely new instrument(s).

Conclusion

An integrated and cyclical framework for quantitative measurement development, validation, and revision has been introduced through extending and adapting the work of Adcock and Collier (2001). This incorporates both qualitative and quantitative methods. It is hoped that this offers a systematic as well as useful practical guide to scholars interested in quantitative measurement development, validation, and revision.

This article also contributes toward mixed methods research literature in a number of ways. It emphasizes the complementarity of qualitative and quantitative methods, and in working from within a constructionist paradigm, supports the view that various paradigms may be applied in the practice of mixed methods research. Paradigmatic choice, in this instance, was determined by an "indigenous" or "particularizing" approach to measurement that recognizes basic conceptual differences may exist between cultural groups in understanding a phenomenon. The philosophical underpinnings of theory would therefore seem important in guiding choice of paradigm rather than a "one size fits all" philosophy of science existing for mixed methods research. Furthermore, mixed methods research is especially beneficial in work that adopts a constructionist outlook, in that it fosters diversity of perspective and seeks to redress some of the power imbalances between the researcher and the researched. The latter is particularly characteristic of quantitative measurement research. In this article, "small-group descriptive" qualitative data and "large-group normative" quantitative data facilitated thorough exploration into varied participant perspectives within and between cultural groups. And last, this article suggests an effective strategy for merging qualitative and quantitative findings in measurement revision through a three-step procedure. Quantitative data are considered before qualitative data. Yet qualitative data are not considered merely supplementary in that they provide unique as well as supplementary information concerning possible revision. This lends some support for the notion that equal status designs are possible in terms of the sum of their individual contributions.

The case study presented in this article only offers an example of how the framework may be applied. It is hoped that it remains flexible enough to be applied to a range of different cases.

Author's Note

Note on terminology: This article makes use of social categories, including those of gender and race. These are understood as social constructs and not essential to individuals. Category descriptors reflect their current construction in South African society. For example, in the case of race, these include Black African, Colored, and White.

Acknowledgments

Special thanks go to Bongaz Maku and Keith Ruiters for their assistance in focus group facilitation, transcription and translation; Maarten Bazuin, Francois Luyt, Margot Luyt and Ntutuzelo Tsotsi for their help in translation; Michael Munsie, Dermot O'Grady, Ntutuzelo Tsotsi and Gareth Watkins for their contribution toward the distribution and collection of questionnaires; Jenny Luyt for her encouragement; and all the men who agreed to take part in this research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article:
Research funds were made available through the Commonwealth Commission and Universities, UK.

References

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review, 95*, 529-546.

Allen, J., & Walsh, J. A. (2000). A construct-based approach to equivalence: Methodologies for cross-cultural/multicultural personality assessment research. In R. H. Dana (Ed.), *Handbook of cross-cultural and multicultural personality assessment* (pp. 63-85). Mahwah, NJ: Lawrence Erlbaum.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*, 33-40.

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

Beere, C. A. (1990). *Gender roles: A handbook of tests and measures*. New York, NY: Greenwood Press.

Ben-Porath, Y. S. (1990). Cross-cultural assessment of personality: The case for replicatory factor analysis. In N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (pp. 27-48). Hillsdale, NJ: Lawrence Erlbaum.

Bertrand, J. T., Brown, J. E., & Ward, V. M. (1992). Techniques for analysing focus groups data. *Evaluation Review, 16*, 198-209.

Brannen, J. (1992). Combining qualitative and quantitative approaches: An overview. In J. Brannen (Ed.), *Mixing methods: Qualitative and quantitative research* (pp. 3-37). Aldershot, England: Avebury.

Brannen, J. (2009). Mixed methods for novice researchers: Reflections and themes. *International Journal of Multiple Research Approaches, 3*, 8-12.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77-101.

Brislin, R. W. (2000). Back-translation. In A. E. Kazdin (Ed.), *Encyclopedia of psychology* (pp. 359-360). Washington, DC: American Psychological Association.

Campbell, D. T., & Fiske, D. A. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Caracelli, V. J., & Greene, J. C. (1993). Data analysis strategies for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 15*, 195-207.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. London, England: SAGE.

Collins, K. M. T., Onwuegbuzie, A. J., & Sutton, I. L. (2006). A model incorporating the rationale and purpose for conducting mixed methods research in special education and beyond. *Learning Disabilities: A Contemporary Journal, 4*, 67-100.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: SAGE.

Denzin, N. K. (1970). *The research act in sociology: The theoretical introduction to sociological methods*. London, England: Butterworth.

Denzin, N. K. (1978). The logic of naturalistic inquiry. In N. K. Denzin (Ed.), *Sociological methods: A sourcebook* (pp. 6-29). New York, NY: McGraw-Hill.

Floyd, F. J., & Widaman, K. E. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*, 286-299.

Gergen, K. J. (1985). The social constructionist movement in modern psychology. *American Psychologist, 40*, 266-275.

Gergen, K. J. (2005). *An invitation to social construction*. London, England: SAGE.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*, 255-274.

Griffin, C., & Phoenix, A. (1994). The relationship between qualitative and quantitative research: Lessons from feminist psychology. *Journal of Community & Applied Social Psychology, 4*, 287-298.

Hammersley, M. (1992). Deconstructing the qualitative-quantitative divide. In J. Brannen (Ed.), *Mixing methods: Qualitative and quantitative research* (pp. 39-55). Aldershot, England: Avebury.

Hoffman, R. M. (2001). The measurement of masculinity and femininity: Historical perspective and implications for counselling. *Journal of Counseling & Development, 79*, 472-485.

Hogan, T. P. (2007). *Psychological testing: A practical introduction*. New York, NY: John Wiley.

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*, 802-812.

Jick, T. D. (1983). Mixing qualitative and quantitative methods: Triangulation in action. In J. Van Maanen (Ed.), *Qualitative methodology* (pp. 135-148). Beverley Hills, CA: SAGE.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112-133.

Kelle, U., & Laurie, H. (1995). Computer use in qualitative research and issues of validity. In U. Kelle (Ed.), *Computer-aided qualitative data analysis* (pp. 19-95). London, England: SAGE.

Levant, R. F., Hirsch, L. S., Celentano, E., Cozza, T. M., Hill, S., MacEachern, M., . . . Schnedeker, J. (1992). The male role: An investigation of contemporary norms. *Journal of Mental Health Counseling, 14*, 325-337.

Luyt, R. (2003). Rhetorical representations of masculinities in South Africa: Moving towards a material-discursive understanding of men. *Journal of Community & Applied Social Psychology, 13*, 46-69.

Luyt, R. (2005). The Male Attitude Norms Inventory-II: A measure of masculinity ideology in South Africa. *Men and Masculinities, 8*, 208-229.

Luyt, R. (in press). Constructing hegemonic masculinities in South Africa: The discourse and rhetoric of heteronormativity. *Gender and Language*.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Morgan, D. L. (2007). Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research, 1*(1), 48-76.

Morgan, G. A., Gliner, J. A., & Harmon, R. J. (2001). Measurement validity. *Journal of American Academy of Child & Adolescent Psychiatry*, *40*, 729-731.

Nagy Hesse-Biber, C. (2010). *Mixed methods research: Merging theory with practice*. London, England: Guilford Press.

Nassar-McMillan, S. C., & Borders, L. D. (2002). Use of focus groups in survey item development. *The Qualitative Report*, *7*. Retrieved from http://www.nova.edu/ssss/QR/QR7-1/nassar.html

Neuman, W. L. (1997). *Social research methods: Qualitative and quantitative approaches*. Boston, MA: Allyn & Bacon.

O'Brien, K. (1993). Improving survey questionnaires through focus groups. In D. L. Morgan (Ed.), *Successful focus groups: Advancing the state of the art* (pp. 105-117). Newbury Park, CA: SAGE.

Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research*, *4*(1), 56-78.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. London, England: SAGE.

Ponterotto, J. G., & Grieger, I. (1999). Merging qualitative and quantitative perspectives in a research identity. In M. Kopala & L. A. Suzuki (Eds.), *Using qualitative methods in psychology* (pp. 49-62). London, England: SAGE.

Rust, J., & Golombok, S. (2008). *Modern psychometrics: The science of psychological assessment*. London, England: Routledge.

Tashakkori, A., & Teddlie, C. (1998). Introduction to mixed method and mixed model studies in the social and behavioural sciences. In A. Tashakkori & C. Teddlie (Eds.), *Mixed methodology: Combining qualitative and quantitative approaches* (pp. 3-19). Thousand Oaks, CA: SAGE.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social & behavioral research*. Thousand Oaks, CA: SAGE.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. London, England: SAGE.

Thompson, E. H., & Pleck, J. H. (1995). Masculinity ideologies: A review of research instrumentation on men and masculinities. In R. F. Levant & W. S. Pollack (Eds.), *A new psychology of men* (pp. 129-163). New York, NY: Basic Books.

Thompson, E. H., Pleck, J. H., & Ferrera, D. L. (1992). Men and masculinities: Scales for masculinity ideology and masculinity-related constructs. *Sex Roles*, *27*, 573-607.

Tran, T. V. (2009). *Developing cross-cultural measurement*. Oxford, England: Oxford University Press.

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-63). Mahwah, NJ: Lawrence Erlbaum.

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée*, *54*, 119-135.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago, IL: Rand McNally.

Wolff, B., Knodel, J., & Sittitrai, W. (1993). Focus groups and surveys as complementary research methods. In D. L. Morgan (Ed.), *Successful focus groups: Advancing the state of the art* (pp. 118-136). Thousand Oaks, CA: SAGE.

Yun, J., & Ulrich, D. A. (2002). Estimating measurement validity: A tutorial. *Adapted Physical Activity Quarterly*, *19*, 32-47.