Psychological Injury and Law (2019) 12:191–203
https://doi.org/10.1007/s12207-019-09362-w

Translated Measures in Forensic Evaluations with Specific Applications to Feigned Mental Disorders

Richard Rogers[1] · John W. Donnelly II[1] · Amor A. Correa[2]

Received: 21 July 2019 / Accepted: 30 September 2019 / Published online: 17 October 2019

Abstract

Forensic assessments continue to grow exponentially from international and transcultural perspectives. As a result, psychological measures are increasingly translated and adapted from their original (source) language to targeted languages. This article begins with broad conceptual issues before proceeding to specific applications. It examines the pitfalls inherent in imposed etic and Western, educated, industrialized, rich, and democratic (WEIRD)-centric adaptation practices. Recommended guidelines for translating and validating tests are discussed, including those promulgated by the International Test Commission (ITC) and the World Health Organization (WHO). The article then focuses on feigning measures—critically important to forensic evaluations—for differentiating between malingered and genuine mental disorders. Finally, feigning research on translated measures of the MMPI-2 and MMPI-2-RF are featured selectively because of their breadth and depth both in general adaptations and their specific attention to feigned mental disorders.

Keywords Translations · International test commission · Imposed etic · WEIRD · Feigning · Simulation designs · MMPI-2 · MMPI-2-RF

International contributions to forensic psychology and psychiatry have only gained prominence in the last three decades. In 1979, David Weisstub, a distinguished Canadian professor of law, launched the *International Journal of Law and Psychiatry* and subsequently played a vital role in the founding the International Academy of Law and Mental Health (IALMH). Still in these early years, Saleem Shah—in his pivotal roles at NIMH—pressed for international issues relating to forensic mental health, including human-rights issues in both China and Russia (Roth, 1995). In reviewing the last 30 years, the international landscape of forensic mental health has grown dramatically. A PsychINFO search (i.e., search terms of "international," "forensic," and either "psychology" or "psychiatry") focusing solely on academic journals provides a mere glimpse at its meteoric growth: 515 entries by 1998, 1360 by 2008, and 3148 by 2018.

✉ Richard Rogers
Richard.Rogers@unt.edu

[1] Department of Psychology, University of North Texas, 1155 Union Circle #311280, Denton, TX 76203-5017, USA

[2] Federal Medical Center: Carswell, Fort Worth, USA

Forensic assessments certainly should not "borrow" psychological measures and trustingly assume that their validity remains intact despite being translated and adapted to very different nationalities and cultures. This point is especially true in forensic contexts because wrong, yet highly consequential, decisions may rely on inaccurate conclusions based on invalid tests. As an instructive illustration, Kelley and her colleagues (Kelley et al., 2018) evaluated the Triarchic Assessment Procedure for Inconsistent Responding (TAPIR) on two forensic samples using the Triarchic Psychopathy Measure (TriPM; Patrick, 2010). When compared to purely random responses (see their Table 5), the false-negative rates were highly problematic for the Dutch (32.5%), yet comparatively good for Swedish (16.5%) detainees. Moreover, the TAPIR varied remarkably in its associations with the TriPM Disinhibition scale from negligibly negative (− .03) for Dutch to moderately strong (.41) for Swedish forensic samples. Are these strong discrepancies resulting from substantive differences in translations and/or the influential effects of culture?[1]

The current article is organized into four major sections that begins with broad conceptual issues and test standards, and

[1] One third of the Dutch sample is composed of persons from outside of Western Europe including Middle and South American, Southern African, and Southern European. Inexplicably, no breakdown by nationality is provided for the Swedish sample.

then illustrates their specific forensic applications. The first section critically addresses two formidable barriers to the effective translation and adaptation of forensically relevant instruments. These barriers consist of (a) the effects of *imposed etics* on tests and their interpretations (Dana, 1993, 2005), and (b) WEIRD (i.e., Western, educated, industrialized, rich, and democratic; Henrich, Heine, & Norenzayan, 2010) sources of knowledge and methodology. The second section carefully examines recommended practices for adaptation of psychological measures to different languages and cultures, including guidelines from two international associations. The third major section provides a methodological framework with reference to feigned mental disorders, a highly relevant determination in forensic assessments. For the fourth section, feigning indicators of the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) and MMPI-2-RF (Ben-Porath & Tellegen, 2008; Tellegen & Ben-Porath, 2011) were addressed selectively, because they constituted the most widely translated and internationally applied multiscale inventories. Finally, the article ends with a brief section of concluding remarks.

Two Barriers to Effective Translation and Adaptation

Imposed Etic

The search for universal or nearly universal principles represents a foundational goal for all sciences. In this regard, the etic approach seeks to find commonalities across cultures and nationalities, which provide an empirical basis for broadly applicable knowledge. The etic approach is frequently combined with emic or culturally specific methods (Cheung, 2004; Zeinoun, Daouk-Öyry, Choueiri, & van de Vijver, 2017). In contrast, "imposed etic" represents researchers' misguided efforts to extrapolate expansively from their own reference point to other cultures. Operationally, this approach has been referred to as the "transport and test function" (Cheung, 2004, p. 179). Regarding imposed etic, Dana (1984, p. 48) offered a withering criticism of such efforts that "distort, caricature, and pathologize persons who are culturally different from ourselves." Imposed etic has also predominated vocational psychology (Leong & Pearce, 2011). In a systematic review spanning the first 96 issues of the *Journal of Cross-Cultural Psychology*, Öngel and Smith (1994, p. 25) found that authors were predominantly US researchers and reached a very troubling conclusion: "Studies (93%) continue to use designs that are predominantly 'imposed etic.'" Unfortunately, we have been unable to find any more recent systematic reviews of imposed etic.

An anonymous reviewer raised a thoughtful comment that the only systematic survey had occurred 25 years ago, raising the possibility that issues with imposed etic may have improved. Clearly, we have no empirical data to either support or question this possibility. However, a recent Google Scholar search (September 25, 2019) suggested that imposed etic remains a salient issue with approximately 170 publications since 2015. In a large-scale study of multi-ethnic health surveys, Pasick, Stewart, Bird, and D'Onofrio (2001, p. 228) appeared to have identified imposed-etics issues: (a) "the question did not allow for cultural differences in meaning" and (b) "the concept or wording was culturally inappropriate."

Cross-cultural researchers have worked to integrate etic and emic (i.e., culturally specific) methodologies. In laying the groundwork, He and van de Vijver (2017) identified major sources of bias that must be addressed in all cross-cultural research. Besides construct biases (i.e., cultural differences in concepts), studies should also consider (a) sample biases (i.e., non-comparable groups), (b) instrument bias (e.g., unfamiliarity with test formats), and administration bias (e.g., unfamiliarity with computerized formats). With these biases in mind, a series of adaptations may take into account differences in language, concepts, culture, and measurement (van de Vijver & He, 2017). Issues of cross-cultural biases should be systematically investigated as potential markers for imposed etic.

WEIRD

Henrich and his colleagues (Henrich et al., 2010) formulated the construct of WEIRD as a non-pejorative acronym that addresses the narrowness of our knowledge base and often has untested assumptions about its transcultural applicability. Citing Arnett (2008), psychological research from 2003 to 2007 utilized samples that "largely reflect the country of residence of the authors, as 73% of first authors were at American universities, and 99% were at universities in Western countries" (Henrich et al., 2010, p. 63). As noted in a commentary, Bennis and Medin (2010, p. 85) described this pattern as "home-field disadvantage" in overgeneralizing from one's own culture. Complementing this perspective, Fessler (2010) argued that cultural congruence between researchers and participants may run the very real risk of overlooking potentially important findings that are not particularly salient in the studied culture.

WEIRD becomes a conceptual framework that underscores the limits of generalizability for translated and adapted psychological measures. Clearly, the involvement of native-speaking researchers from the country and culture of interest is vital to the process. Because psychologists and other social-science researchers may be heavily influenced—even several generations later—by WEIRD trained academics, consideration should be given to active involvement of non-WEIRD community members. Such proactive engagement and true

collaboration becomes an essential requirement when involving indigenous populations (e.g., Black, Toombs, & Kisely, 2018; Nwoye, 2015). Standards and concomitant research (see, e.g., Fitts et al., 2015; National Health and Medical Research Council, 2018) with Australian first peoples may provide a useful template on ethical practices that are broadly applicable to other cultures.

Insufficient cross-cultural competence (i.e., 3C) on the part of WEIRD-trained researchers may limit the success of adaptation processes. Gathering feedback on researchers' abilities to work with diverse populations via validated 3C instruments (Matsumoto & Hwang, 2013) may provide insight as to participants' perceptions of the test development process. More intensive collaborative approaches often outline methods for gathering input from culturally diverse participants using focus groups (Alegria et al., 2004) or cognitive pre-testing (ITC, 2016; UMN, 2011; WHO, undated). Given research demonstrates that self-report measures of 3C are often uncorrelated with actual abilities and largely reflect social desirability (Constantine & Ladany, 2000), observer-reported instruments offer a distinct advantage. The Multicultural Personality Questionnaire (MPQ; van der Zee & van Oudenhoven, 2000) represents one such measure. MPQ observer ratings have been shown to provide a reliable and valid assessment of behavioral effectiveness in multicultural contexts (Matsumoto & Hwang, 2013; van der Zee & van Oudenhoven, 2001).

As a side note, the independent development of emic or culture-specific instruments may mostly circumvent the negative impacts of "imposed etic" translations (Cheung, Cheung, Wada, & Zhang, 2003). As observed by Cheung et al. (2001), this approach is particularly advantageous when measuring constructs found in specific cultures (e.g., *harmony* and *face* in Asian cultures). For example, the Chinese Personality Assessment Inventory (CPAI; Cheung et al., 1996) represents a valuable emic instrument with applicability to forensic contexts (Cheung, Kwong, & Zhang, 2003).

In real-world applications, practitioners may wish to systematically consider the relative role of WEIRD constructs in test adaptations in determining whether a particular measure has practical usefulness. While recognizing the substantial interrelationships among the five WEIRD constructs, clinicians may wish to weigh their combined limits on cross-cultural generalizability, particularly as it applies to normative groups and interpretative cut scores. For forensically relevant instruments in particular, the risk of highly consequential misclassifications is likely to outweigh all other clinical considerations. For this reason, both translation and adaptation initiatives must carefully consider what combination of emic and etic constructs is most relevant to forensic assessments in terms of validity, utility, and cultural fairness.

Recommended Practices for Translations, Adaptations, and Validations

Different professional organizations have promulgated recommended practices and standards for translations and validations of psychological measures (see Table 1). First and foremost, the International Test Commission (ITC; Hambleton, 1996) was formed more than two decades ago with input and membership from five international organizations including the disciplines of psychology, education, and language. Now in its second edition, the ITC (2016) provides the most comprehensive guidelines involving all aspects of test development, validation, and documentation. In addition, the World Health Organization (WHO) and National Institute of Health (NIH), via their programmatic research on National Latino and Asian American Study (NLAAS), have also developed their own recommendations. To broaden this analysis, we reviewed major test publishers for their guidance in test translations. Pearson Assessments, a leader in adapted measures, publishes MMPI-2 and MMPI-2-RF translations that follow the University of Minnesota Press (UMN, 2011) Guidelines for Developing Translations. Therefore, the UMN guidelines were added to this review.

Prior to initiating the translational process, researchers should obtain formal permission from the copyright holder and authors of the intended measure to avoid any legal or ethical issues (ITC, 2016). Beyond the permission itself, publishing companies understandably have a vested interest in ensuring quality products that will enhance and not diminish their professional reputation. Therefore, formal agreements typically specify objectives, procedures, and products that will occur as part of the agreed-upon translation.

The \ITC (2016) prudently recommends detailed planning prior to the initiation of the translation. At the most basic level, researchers should consider the worthiness of investing in this time-intensive endeavor rather than pursuing independent development of a culturally relevant instrument in the selected language (see, e.g., Cheung et al., 2003). If the decision is affirmative, then content and cultural experts should actively collaborate with the related goals of anticipating potential issues and proactively seeking to minimize them (ITC, 2016).

The primary objective of the following seven subsections is to succinctly summarize key issues in translating and adapting psychological measures. It should allow practitioners—after reviewing a measures' translation and adaptation procedures—to make informed decisions about its cultural applicability and clinical validity.

Project Staff

The selection of project staff must take into account their language skills, cultural competence, and assessment backgrounds. ITC (2016) stresses the importance of including

Table 1 Standards for adaptations of psychological measures to different languages and cultures

Criterion	Sources			
	ITC	WHO	NLAAS	UMN
Project staff				
→ Team member with primary language/culture of intended translation	✓	✓		✓
→ Team member with expertise in psychological assessment or psychometrics	✓	✓		✓
Etic or emic emphasis				
→More etic with diverse ethnic backgrounds for diverse applications			✓	
→ More emic with focused ethnic backgrounds for focused applications	✓	✓		✓
Forward translation				
→ Single or unspecified		✓	✓	
→ Multiple	✓			✓
→ Mental health professional with an assessment training	✓	✓		✓
Independent back-translation				
→ Entirely separate from the forward translation	✓	✓	✓	✓
→ No working knowledge of the original measure	✓			
Further review and revisions				
→ Input for native speakers on the meaning of items	✓	✓	✓	✓
→ Expert panels for reviews and revisions	✓	✓	✓	✓

ITC = International Test Commission; *WHO* = World Health Organization; *NLAAS* = National Latino and Asian American Study; *UMN* = University of Minnesota; *NA* = not applicable

project staff who know both the language and the culture. Employing the term "localization" emphasizes culture-specific adaptations. In contrast, some NLAAS research (Alegria et al., 2004) utilized an international team of bilingual investigators in seeking to establish a "common" translation. Unfortunately, the culture, nationality, and language of project staff are sometimes unspecified. Such omissions militate against an informed appraisal of their backgrounds and expertise.

Some members of the project staff should possess in-depth knowledge of psychological testing as well as specific competency with the measure to be translated. UMN (2011) expressed a preference for a bilingual academic psychologist assuming a leadership position. In addition, ITC (2016, p. 10) provides an operational definition, "experts with respect to the construct measured, and who are familiar with the cultural groups being tested, should be recruited to evaluate the legitimacy of the construct measured in each of the cultural/linguistic groups." From the perspective of NLAAS (Alegria et al., 2004), the primary emphasis involved the formalized input from an ethnically diverse panel of representative experts.

Etic or Emic Emphasis

The recommended guidelines should not be categorized as completely etic or emic. Thinking dimensionally, the NLAAS (Alegria et al., 2004) has a more etic emphasis than the other professional guidelines. These investigators examined the feasibility of test items across four Latino groups: Mexican, Puerto Rican, Cuban, and Other Latino. They sought items relevant across cultural groups (etic emphasis) rather than being culturally specific (emic emphasis). Through their concerted efforts, these researchers established a common-language translation with only a few exceptions that relied on culturally specific vocabulary. While reducing many within-cultural differences, Alegria et al. (2004) still encountered many enduring transcultural challenges between mainstream US culture and the selected Latino cultures. As a salient example, linguistic equivalence could not be adequately established with an English-language NLAAS measure of personality.

The initial translations—sometimes referred to as the "forward translations"—represent the first crucial step in adapting a psychological measure to a new language. Although researchers sometimes use a single translator (Peña, 2007), recommended practices generally rely on multiple translators, either from (a) the same culture and country (UMN, 2011) or (b) representing diverse cultures with the goal of eliminating "non-consensus" words (Alegria et al., 2004). Surprisingly, WHO (undated, para. 3) suggested "One translator, preferably a health professional, familiar with terminology of the area covered by the instrument." Despite this exception, the general consensus favors the additional rigor provided when multiple professionals are independently engaged in the forward translation.

The approval of the forward translation varies substantially across recommended guidelines. For UMN (2011, para. 2), the translators simply compare their independently translated versions "negotiating differences in the translation of items to obtain the most equivalent item." More typically, the initial translations are reviewed by an expert panel, functioning independently of the measure's forward translators.[2] The ITC (2016) cited with approval the systematic approach by Hambleton and Zenisky (2011, pp. 49–52) to evaluate item comparability. Items must have the same or highly similar meanings and be written at similar levels of difficulty. They must also be checked for (a) any alterations (e.g., additions, substitutions, or omissions) and (b) the absence of metaphors, idioms, or colloquialisms that may alter the intended meaning. Importantly, Hambleton and Zenisky (2011) provide empirical justifications for their systematic approach, including alterations (Collazo, 2005; Gierl & Khaliq, 2001), and distortions in meaning (Collazo, 2005; Van Eeden & Mantsha, 2007).

The back-translation from the target language to the source language is performed by one or more translators. According to WHO (undated), the back-translator should have the source language as his or her native language. Most recommended guidelines emphasize the importance of back-translations being conducted by different translators, completely independent of the forward translations. Except for WHO (undated), recommended guidelines have generally omitted another form on non-independence. This source of non-independence is italicized to emphasize its importance: *Back-translators should have no knowledge of the translated measure in its source language*. Clearly, knowledge of the items in its source language is likely to strongly inflate the back-translation agreement because of "insider knowledge." This point is explicitly recognized by Kiing, Rajgor, and Toh (2016, p. 1112, Table 1) in requiring a "bilingual translator not familiar with the tool, i.e., naïve to the questionnaire."[3]

Further Review and Revisions

Pilot research may address how culturally relevant populations understand the meaning of the test items and test instructions. As an example of the former, simple queries can be implemented, such as the following, "Tell us in your own words, what this question is asking." WHO (undated, para. 10) offered specific recommendations for seeking test-taker insights: "what they thought the question was asking, whether they could repeat the question in their own words, what came to their mind when they heard a particular phrase or term." On this point, ITC (2016) astutely observed that test-takers' comments about the measure may be more valuable than their actual responses to the measure.

ITC (2016) recommended checking item-level comparability using culturally relevant bilingual respondents. Unlike formal testing of linguistic equivalence, this analysis would be focused on item revisions, specifically the early identification of potentially problematic translated items so they can be subsequently revised. Although omitted from other recommended guidelines, this step for improving item-level comparability appears far superior to waiting until formal testing has been conducted and then retroactively attempting to revise or even exclude problematic items from the translated version.

Expert panels for reviewing translated measures vary substantially across standards in their level of rigor and detail. For instance, UMN (2011; para. 4) simply calls a review by the university's Language Service in "assessing equivalence, attention is paid to vocabulary, idiom, syntax, and tone." If this service is located primarily in the USA, WEIRD influences may potentially distort their results. Both ITC (2016) and NLAAS (Alegria et al., 2004) guidelines emphasize culturally competent experts. In particular, NLAAS uses multiple focus groups from each country or region to address issues of translation and cultural relevance. Presumably independent of each other, the use of the multiple groups provides a highly valuable within-culture cross-check for ensuring the applicability of the translated measure to the targeted population.

Linguistic and Cultural Equivalence

Both UNM (2011) and the ITC (2016) recommended the testing of linguistic and cultural equivalence with culturally relevant bilingual samples. UNM (2011, para. 5) authorized the use of bilingual participants, "who are fluent in both languages and familiar with both cultures." ITC (2016) advised the evaluation of the test instructions as well as test items with bilingual participants. To test for equivalence across language and culture, this commission suggested IRT-based differential item functioning (DIF) analyses to formally evaluate the comparability of source and target versions. As readily acknowledged by the ITC, however, such analyses involve very large samples, thus limiting the practical applicability of DIF. At a more general level, scales may be directly compared with respect to their internal consistencies (Alegria et al., 2004; also, see next section).

Internal Consistencies and Reliability

ITC (2016) raised strong concerns about any substantial decreases in test score reliability from the source measure to the

[2] Unfortunately, WHO may unintentionally introduce unidentified biases by typically including the original translator, who presumably is expected to be completely objective in his or her review of their own work.

[3] These practices are recommended by the American Association of Orthopedic Surgeons (AAOS) with the unequivocal requirement that back translators are "totally blind to the original version" (Beaton et al., 2000, p. 3188).

targeted translation. For instance, a much lower alpha would heighten concerns regarding the translation and its cultural applicability. Alegria et al. (2004) recommended systematic comparisons of alpha coefficients across different language versions. For their own investigation, internal consistencies were highly comparable with very similar alphas for the source version ($M = .82$) and the targeted language ($M = .83$). Comparisons of alphas and inter-item correlations (Beaton, Bombardier, Guillemin, & Ferraz, 2000) should be an essential and easily implemented component for test translations.

Reliability varies with test formats with options of split-half, interrater, and test-retest reliabilities (ITC, 2016). When feasible, reliability coefficients should be augmented by standard errors of measurement (SEM; UNM, 2011) and 95% confidence intervals. Unfortunately, most of the recommended guidelines provide very little specific information concerning the types of reliability. From a test-retest perspective, for example, the concordance rates for MMPI-2 two-point codes between two administrations of the targeted version would be highly valued by practitioners in bolstering their conclusions. As an instructive analogue, Edwards, Morrison, and Weissman (1993) found only a modest concordance of 58% between the MMPI and MMPI-2 codes for the same outpatients.

Validity

The highly respected *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 2014, Standard 7.6, p. 127) require that translated measures be independently evaluated in the target language. It clearly specifies the following: "Whenever tests are translated from one language to a second language, evidence of the validity, reliability/precision, and comparability of scores on the different versions of the tests should be collected and reported." Although achieving similar scores across different languages may have some merit, the ITC (2016, p. 22) has firmly established the crucial priority: "careful examination of the validity of the test in the second language group is essential." Moreover, whenever feasible, validity should be examined separately for each different culture or nationality (see AERA/APA/NCME, 2014, Standard 3.12, p. 69).

The ITC (2016) provided the most comprehensive guidelines for the validation of translated measures. In addition to convergent and discriminant validity, their test guidelines address construct validity in detail. Confirmatory factor analysis—as well as other statistical methods—is recommended for directly evaluating the equivalence of source and targeted language versions.

Clinical Applications

When applicable, appropriate norms must be established for the translated version using culturally relevant samples. While it is conceivable that strong empirical evidence could be accumulated to support the original norms from the source language, the establishment of new norms appears to be the more prudent course, based on the intended populations (ITC, 2016). On this point, UMN (2011, para. 6) recommended large numbers (i.e., ≥ 350 persons for each gender) be collected that are demographically representative, which are augmented with a "demographically representative clinical sample of 100 men and 100 women."

Norms can only be interpreted in the context of external validation. Consider for a moment a new measure of academic achievement. Irrespective of names given to its scales, what possible value can be derived from an unvalidated normative score? This crucial point is sometimes entirely overlooked with translated multiscale inventories. Stated succinctly, the ITC (2016) recommended guidelines coupled (a) norms for translated measures with (b) research on their reliability and validity. In addition, the AERA/APA/NCME (2014) clarified that interpretations for different languages should be formally evaluated and ensure that score interpretations from the two versions have comparable validity for their intended uses.

In clinical and forensic practice, specific interpretations are typically provided in measures of psychopathology and response styles when certain criteria are reached or exceeded. With the Personality Assessment Inventory (PAI; Morey, 2007), for example, interpretations and correlates are generally provided for "moderate" (T scores from 60 to 69) as well as "clinical" (T scores ≥ 70) elevations (see, e.g., Rogers, Williams, Winningham, & Sharf, 2018). Criteria for interpretations are usually expressed in terms of elevations or cut scores. With reference to forensic relevant instruments, cut scores are frequently used with feigning measures in providing dichotomous or polychotomous classifications.

Recommended guidelines for translated measures have yet to address the independent validation of elevations and cut scores. In rendering clinical and forensic conclusions, the empirically established accuracy of specified elevations and cut scores may be viewed as an essential component in validating psychological measures for their real-world applications. We propose that professional organizations and commissions consider the following addition, placed in italics for emphasis: *Elevations and cut scores for each translated measure must be externally validated before they may be ethically implemented in professional practice*. For elevations, empirical correlates—completely independent of the translated measure—should be examined by culture, nationality (e.g., Moultrie & Engel, 2017), and gender (e.g., Rogers et al., 2018). For cut scores, classification accuracy (e.g., sensitivity and specificity) should be formally evaluated by culture

(Nijdam-Jones & Rosenfeld, 2017). In stressing its importance, this recommended guideline is included in Table 2 separately for elevations and cut scores.

This section outlined the important issues for the translation and validation of psychological measures. In focusing of feigned measures, the next major section presents an easily understood overview of research designs for response styles. It sets the stage for the final major section which briefly illustrates issues with adaptations in the context of feigned measures.

Three Methodological Considerations for Measures of Feigned Mental Disorders

Research on response styles varies substantially in terms of their design and level of methodological rigor. Besides reviewing the quality of particular translations, practitioners will also need to critically evaluate the quality of feigning research before implementing any translated measure in their forensic practice. To simplify this process, Table 3 provides a useful checklist that can be applied to the targeted language versions.

The two strongest designs for studies of feigning involve known-group comparisons (KGCs) and well-executed simulation (SIM) designs (Rogers & Gillard, 2011). The most notable strength of KGC designs is their external validity, utilizing actual examinees in their real-world settings, such as forensic evaluations. In comparison, well-designed simulation studies excel in internal validity using standardized methods, such as random assignments, identical scenarios, and formalized manipulation checks. The strongest validation of feigning measures includes convergent data from both KGC and SIM models.

Criterion-Based Validations

The mere presence of any external criterion most often falls far short of a KGC design. Rogers (2018b) carefully distinguished KGC from partial criterion designs (PCDs). For establishing *known groups*, two crucial methodological issues involve minimizing classification errors and measurement errors. Groups cannot be correctly determined if the criterion measure lacks accuracy. As a simple analogy, fevers cannot be established, if the thermometer varies by ± 5 degrees centigrade. Although this analogy might be summarily dismissed as inapplicable, recent research (Tylicki et al., 2018) apparently failed to consider such variability in using the MMPI-2-RF F-r < 80T as a cut score for genuine responding. With a standard error of measurement of 10T, practitioners should be very concerned about the accuracy of this classification when the 95% confidence interval falls some place between typical responding for clinical groups (i.e., 59T) and an extreme elevation associated with feigning (i.e., 99T).

Recommended guidelines (Rogers, 2018b, p. 602, Table 30.2) are summarized for differentiating KGC and PCD designs:

- KGC "uses the best validated measure," "applies stringent criteria to all involved groups," and "removes an indeterminate group (i.e., too close to call)."
- PCD "uses a moderately validated scale," "applies criteria to only the response style of interest (e.g., malingering)," and "uses all participants."

PCD research with no participants excluded may push the percentage of overall errors over 50% when measurement and classification errors are considered together (e.g., Rogers, Gillard, Wooley, & Ross, 2012). Practitioners may wish to either minimize or entirely avoid conclusions about feigning based on PCD studies when failures exceed successes. In sharp contrast, KGC studies may provide strong evidence of external validation.

Well-Executed SIM Designs

SIM designs have been developed and refined over the last two decades (Rogers, 1997, 2008, 2018b). In most instances, the differences between well and poorly designed SIM studies simply involve systematic attention to detail. By design, SIM research does not exist in the real world in terms of (a) situational stressors, (b) the scenario or setting, (c) the often life-altering stakes of the evaluation, and (d) the devastating consequences if efforts at feigning are detected. The following paragraphs summarize essential components of SIM designs that parallel the checklist of key issues.

Engagement and Investment Studies without manipulation checks have little clinical or forensic relevance (see no. 3 in Table 3). Typically, 5 to 10% of simulators do not accurately recall the instructions or they candidly acknowledge a lack of sustained effort. Others with sufficient abilities may respond carelessly—another compelling example of being uninvested in the study. A 1-min manipulation check may remove a substantial level of irrelevant data from SIM studies. Beyond manipulation checks, what efforts were implemented to motivate participants to take the study seriously? While external incentives may be small, analogous to penny poker, they may still improve performance if simulators are convinced that only "successful" simulators will receive them. As an internal motivation (see no. 4), Rogers (2018b, p. 601) recommended challenging simulators concerning their skills: Are you capable enough to "beat the test?" As a parallel

Table 2 Validation of psychological measures with different languages and cultures

Criterion	Sources			
	ITC	AERA	NLAAS	UMN
Equivalence				
→ Linguistic equivalence: bilingual testing	✓			✓
→ Cultural equivalence: results across cultures	✓	✓	✓	✓
Reliability				
→ Scale homogeneity	✓	✓	✓	✓
→ Test reliability	✓	✓		✓
Validity				
→ Construct validity (e.g., factor analysis)	✓	✓		✓
→ Convergent validity	✓			
→ Discriminant validity	✓			
Standardization				
→ Standardized scores				✓
→ Norms	✓			✓
Clinical applications				
→ Accuracy of elevated scores	✓			✓
→ Accuracy of decision rules (e.g., cut scores)				

ITC = International Test Commission; *AERA* = American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education; *NLAAS* = National Latino and Asian American Study; *UMN* = University of Minnesota; *NA* = not applicable

to real-world applications, simulators should be warned about the consequences of "failed" simulation (no. 5). Such warnings may also include a cautionary statement about the presence of validity indicators designed to detect feigned mental disorders.

Clinical Relevance The weakest SIM designs for feigning rely exclusively on undergraduates without any clinical comparison group and with backgrounds (e.g., education and verbal abilities) generally incompatible with their intended applications. While unacknowledged, such designs should plainly be characterized as an imposed etic—and likely WEIRD—in their assumptions. From a practitioner's perspective, undergraduate-only studies entirely miss the crucial real-world challenge of accurately differentiating between genuine and feigned mental disorders (no. 6). To underscore this obvious point, *SIM studies without relevant clinical comparison groups provide no evidence of discriminant validity.* Without the appropriate clinical comparisons, patients with genuine disorders may well be misclassified as malingerers (e.g., Peters, Jelicic, Moritz, Hauschildt, & Jelinek, 2013).

Simulators should not be asked to feign in a setting or situation that is completely alien to their own life circumstances. To the extent possible, simulators should be able to relate to the feigning scenario and be able to draw on their own experiences (Rogers & Cruise, 1998), such as asking patients with some genuine PTSD symptoms to feign disabling PTSD (Wooley & Rogers, 2015).

Comparability Practitioners may readily identify how the lack of comparability between clinical comparison and feigning groups may militate against any meaningful conclusions. Using Spanish as the translated language, for example, how can results from fully bilingual and presumably acculturated undergraduates from an English-speaking university (see, e.g., Fernandez, Boccaccini, & Noland, 2008) be seen as remotely comparable to Spanish-only patients with mostly traditional values and limited education in any language? Therefore, Table 3 summarizes the essential points of comparison regarding general backgrounds, culture and nationality, and language abilities.

Selective Illustrations of Problematic Issues with the Translated MMPI-2 and MMPI-2-RF

A comprehensive review of all translated feigning measures would require a book-length approach. Instead, this section selectively illustrates methodological considerations arising from feigning measures as they relate to translated versions of the MMPI-2 and MMPI-2-RF. These closely related measures constitute the obvious choice, given (a) the breadth and depth of their translations and (b) the existing body of feigning research on

Table 3 Checklist of key issues for translated feigning measures

Check	Design	Methodological issue
	KGC	1. Known-groups comparison (KGC): uses the best criterion and removes an indeterminate group (i.e., "too close to call")
	PCD	2. Partial criterion design (PCD): typically uses of multiple convergent feigning measures but leaves in the indeterminate group
	SIM	3. Rigor of Simulation (SIM) design: manipulation checks on recall of instructions and rated effort
	SIM	4. Rigor of SIM design: provides simulators with external incentives or internal motivations
	SIM	5. Rigor of SIM design: cautions simulators about real-world consequences of failed feigning
	SIM	6. Rigor of SIM design: uses clinical comparison sample relevant to research goals (e.g., criminal forensic evaluations)
	SIM	7. Rigor of SIM design: provides a clinically relevant scenario that simulators can find relevant
	All	8. Comparability: Similar general backgrounds for clinical comparison and feigning groups
	All	9. Comparability: Similar culture and nationality for clinical comparison and feigning groups
	All	10. Comparability: Similar language proficiencies for clinical comparison and feigning groups

PCD = partial criterion design; *SIM* = simulation design

various translations. On the matter of breadth alone, translations licensed by the University of Minnesota Press include 24 for the MMPI-2 and 12 for the MMPI-2-RF. Using Spanish as an example, three translations are listed on the website (https://www.upress.umn.edu/test-division/translations-permissions/available-translations) as currently available: (a) Spanish for Mexico and Central America; (b) Spanish for Spain, South America, and Central America; and (c) Spanish for the USA. Each Spanish translation contains key language variations that reflect the vernacular of the region and are not effectively interchangeable due to potential differences in word meanings that could compromise construct validity. Using an English example, an "apartment" in American English is referred to as a "flat" in British English and a "unit" in Australian English, while "flat" and "unit."

This section recognizes the current accomplishments with the MMPI-2/MMPI-2-RF and looks forward to future developments. For practitioners interested in a formal meta-analysis, systematic reviews are already underway (Bopp, 2019; see also Aparcero-Suero, 2019). The first subsection focuses on validity scales for the translated measures.

Validity Scales

UMN (2011) Guidelines for Developing Translations (https://www.upress.umn.edu/test-division/translations-permissions/GUIDELINES) clearly acknowledged that validity indicators often require adaptations for translated versions. For example, each item pair on inconsistency scales needs to be evaluated individually with ineffective pairs being deleted and replaced, whenever possible. Similarly, item frequencies are examined individually for the infrequency scales with the same process of deletion and replacement. Both approaches utilize both normative and clinical samples. In our selected review of the MMPI-2/MMPI-2-RF, however, none of the studies included such critically important details.

Criterion-Based Studies

Criterion-based research has rarely been undertaken with MMPI-2 and MMPI-2-RF translations and would be mostly classified as partial criterion design or PCD (Rogers, 2018b). For example, De Marchi and Balboni (2018) grouped male inmates as "feigning" if the onset of their serious mental disorders reportedly occurred after imprisonment with scores lower than expected on the Symptom Check List-90-Revised (SCL-90-R) in reporting psychopathology. Although a sudden onset would be potentially unexpected for many mental disorders (Rogers, 1984), it has not been proven to be a valid indicator of feigned mental disorders. In addition, Chang, Tam, and Chiang (2017) utilized a component of PCD in testing whether undergraduate simulators would likely be classified as feigning using the Structured Interview of Reported Symptoms–Second Edition (SIRS-2; Rogers, Sewell, & Gillard, 2010) as the external criterion. As briefly illustrated by these two studies, PCD designs could be improved for feigning research with the MMPI-2/MMPI-2-RF.

Fariña, Arce, Vilariño, and Novo (2014) utilized an innovative KGC design with legally adjudicated victims of intimate partner violence (IPV). The evaluations of the adjudicated victims were conducted at a university-based Forensic Psychology Institute, but were apparently not used in any legal proceedings. The study could be further strengthened by the inclusion of diagnostic data and means for MMPI-2 clinical elevations so that its generalizability and applicability to other samples could be evaluated.

Clearly, opportunities abound for KGC and empirically strong PCD studies on translations of the MMPI-2/MMPI-2-RF. With respect to the latter, PCD studies can easily be improved by adding multiple indicators, each with strong empirical support. However, care should be taken that the multiple indicators represent different detection strategies (Rogers, 2018a) and are not redundant (i.e., highly correlated). When

a comprehensive measure of feigned mental disorders, such as the SIRS-2, has been validated in the target language (e.g., Spanish version for the USA), then the valuable KGC design may be straightforwardly implemented, making sure to omit the indeterminate groups via the SIRS-2 decision model (Rogers et al., 2010).

Simulation Studies

For the sake of clarity, MMPI-2/MMPI-2-RF examples follow the same organization as early section, entitled "Well-Executed SIM Designs." As noted previously, the goal is to illustrate strengths and challenges rather than provide a comprehensive review.

Engagement and Investment Hahn (2005) invested simulators in the study (e.g., feigning instructions were repeated twice) and estimated their overall effort via a manipulation check. Giger, Merten, Merckelbach, and Oswald (2010) went a step further in pre-testing participants' recall of their simulation instructions via multiple-choice questions, which was subsequently followed by a manipulation check. In motivating participants to become fully engaged in their simulations, research sometimes stresses the real-world importance of their findings. In studying IPV, Fariña et al. (2014, p. 3) reminded simulators regarding the far-reaching ramifications of faked accusations of partner violence including wrongful convictions as well as "indirect harm and suffering to children." Finally, simulators may be cautioned about feigning in a credible manner to avoid detection via validity indicators. However, care must be taken to write this simply and clearly. Highly complicated warnings, such as a 36-word sentence requiring reading comprehension at a college education, should be avoided (e.g., Kopf, Galić, & Matešić, 2016).

Studies of translated MMPI-2/MMPI-2-RF tend to favor external incentives over internal motivations. For instance, Kopf et al. (2016) provided research credit to college students, indicating this incentive was limited to only successful feigners. In contrast, Arce, Fariña, and Buela (2008, p. 488) provided a substantial financial incentive (i.e., 150 Euros) to "the four best feigners of psychological injury." Future research should definitely consider direct challenges regarding simulators' effectiveness at avoiding detection; such improvements are easily implemented to address internal motivation.

Clinical Relevance Simulation studies without clinical comparison samples should not be used to assess feigning, because the very real risks of false-positives (i.e., misclassifying genuine patients as feigners) cannot be ignored. For the Korean MMPI-2 translation (Hahn, 2005), genuine patients averaged extreme elevations on F ($M = 84.61$), Fb ($M = 89.55$), and even Fp ($M = 93.67$). Unfortunately, a substantial minority of translated studies fail to address this essential component. For authorized translations by UMN (2011), however, a practical solution may exist. As previously mentioned, their standards require a minimum clinical sample of 200. These data could be utilized to establish cut scores for the MMPI-2/MMPI-2-RF, although their applicability may potentially be limited to a specific setting (e.g., outpatients only).

Scenarios vary remarkably on two parameters: (a) their familiarity to simulators and (b) their applicability to clinical and forensic evaluations. Some studies (e.g., Hahn, 2005) are straightforward for simulators with goals of seeking a psychiatric admission and monetary compensation. Sánchez, Ampudia, Jiménez, and Amado (2017, p. 53) offered alternatives: "to avoid a criminal prosecution, to obtain financial compensation, and/or to seek revenge," but unfortunately did not make any post hoc comparisons of whether their results differed significantly by the scenario objectives.

Some other MMPI-2/MMPI-2-RF studies asked simulators to adopt unfamiliar and extreme roles. Chang and her colleagues (Chang, Tam, Shiah, & Chiang, 2017, p. 4) asked simulators—mostly college students—to imagine the very extreme circumstances that "they had killed someone and were eager to escape culpability by feigning psychosis." A companion study (Chang, Tam, Shiah, & Chiang, 2017) further intensified the scenario by specifying murder and the death penalty. As an additional example, Giger et al. (2010) provided a highly imaginative but equally improbable scenario that involved stealing secrets for a profit, physically striking with "great force" a life-sized dummy, and feigning amnesia after being charged with the dummy's homicide.

Practitioners need to review the feigning studies for clinical applicability in their own professional settings. Research with large samples and broadly applicable scenarios (e.g., Sánchez et al., 2017) have a strong empirical basis. Other studies may be either inapplicable or narrowly applicable. For the latter, research may be circumscribed to a specific clinical condition, such as mixed anxiety-depressive disorder (Kopf et al., 2016) or a particular referral issue, such as false allegations of IPV resulting in psychological injury (Fariña et al., 2014).

Comparability The clinical usefulness of simulation studies may be severely constrained when the backgrounds of feigning and clinical comparison samples are markedly different. College simulators are likely to differ substantially from most clinical samples in verbal abilities and education. Other key differences of relevance to translated version may be increased familiarity with idioms and content from the source language. For example, Hahn (2005) found remarkable differences on MMPI-2 feigning scale between college and psychiatric samples under standard instructions. Average raw scores for F, Fb, and Fr ranged from 4.12 to 8.78 in students to 7.51 to 15.69 in patients. On the MMPI-2-RF, a similar finding was observed for F-r (5.25 versus 10.47) but not Fp-r (Sánchez et al., 2017).

New avenues of research should take into account the potential effects of source language and transcultural knowledge in feigning studies with the MMPI-2/MMPI-2-RF. Moreover, understanding and familiarity with mental disorders should be investigated, because they may vary across culture and thus be reflected in both validity and clinical scales.

Concluding Thoughts

The last three decades have documented important advances in translations and adaptations of psychological measures, including forensically relevant instruments. A detailed review of these professional guidelines established common recommendations for the translation and back-translation process. In planning future translations, early preparation with forward-thinking decisions should be strongly underscored at the very beginning in the selecting of measures and professional staff, and addressing possible setbacks. To ensure methodological rigor, two key themes consistently involve (a) diverse perspectives with two or more professionals at each stage of the translation and validation that (b) function independently both within and across stages. To achieve the latter, each professional is involved only once so that their objectivity is unnecessarily jeopardized by others' input.

Standardization can be used to improve methodological rigor and reduce subjectivity. As cogently observed by an astute reviewer, those strengths do not necessarily justify its implementation. Before any adaptation is even considered, issues involving WEIRD and imposed etic must be openly discussed by fully engaging the indigenous community and implementing culture-specific adjustments to assessment methods, when needed.

Examination of cultural influences, while often characterized as "outward looking" (e.g., participants and measures), may also be viewed from a self-reflective perspective. How does the investigator's own cultural background and transcultural awareness potentially affect research? Assessments of cross-cultural competence may range from informal introspection to the self-administration of formal measures (Matsumoto & Hwang, 2013). Within a team of researchers, feedback could voluntarily be sought from colleagues to improve our cross-cultural awareness and humility.

Studies of feigning in both source and target languages may often be easily improved by following recommended guidelines (Rogers, 2018b) including those found in Table 3. It would be very helpful if these guidelines were made readily available to authors and reviewers, asking the former to complete a brief checklist. A similar format might also be implemented for translated measures.

As a final note, forensic evaluations of response styles continue to face formidable challenges in their translations, adaptations, and validations. It may be tempting to compile research conducted in "other" languages and study their apparent similarities. However, seasoned forensic practitioners should easily recognize both the implicitly imposed etic and the methodological weakness of categorically ignoring the important effects of language, culture, and nationality. Looking to the next three decades, it is our hope that collaborative research on feigning and other response styles continue to grow in their depth, sophistication, and cultural sensitivity.

Compliance with Ethical Standards

This article complied with APA ethical standards.

Conflict of Interest The authors declare that they have no conflict of interest.

References

Alegria, M., Vila, D., Woo, M., Canino, G., Takeuchi, D., Vera, M., et al. (2004). Cultural relevance and equivalence in the NLAAS instrument: Integrating etic and emic in the development of cross-cultural measures for a psychiatric epidemiology and services study of Latinos. *International Journal of Methods in Psychiatric Research, 13*(4), 270–288 Retrieved from https://libproxy.library.unt.edu:2147/10.1002/mpr.181. Accessed 6 July 2019

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Aparcero-Suero, M. (2019). *Detecting symptom exaggeration and defensiveness with the MMPI-2: An updated meta-analysis*. Montreal, QC: International Association of Forensic Mental Health Services.

Arce, R., Fariña, F., & Buela, G. (2008). Assessing and detecting the ability to faking psychological injury as a consequence of a motor vehicle accident on the MMPI-2 using mock victims. *Revista Latino-Americana de Psicología, 40*, 485–496.

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *The American Psychologist, 63*(7), 602–614.

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine, 25*(24), 3186–3191.

Bennis, W. M., & Medin, D. L. (2010). Weirdness is in the eye of the beholder. *Behavioral and Brain Sciences, 33*(2–3), 85–86.

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2-RF, Minnesota Multiphasic Personality Inventory-2 Restructured Form: Manual for administration, scoring and interpretation*. Minneapolis, MN: University of Minnesota Press.

Black, E. B., Toombs, M. R., & Kisely, S. (2018). The cultural validity of diagnostic psychiatric measures for indigenous Australians. *Australian Psychologist, 53*(5), 383–393.

Bopp, L. (2019). *Detecting malingering and defensiveness in the translated versions of the MMPI-2 and MMPI-2-RF*. Montreal, QC: International Association of Forensic Mental Health Services.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2: Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

Chang, Y. T., Tam, W. C. C., & Chiang, S. K. (2017). Detection of feigning psychosis with multiscale personality inventories: A

simulation design pilot study in Taiwan. *SAGE Open, 7*(3), 1–9. https://doi.org/10.1177/2158244017734023.

Chang, Y. T., Tam, W. C. C., Shiah, Y. J., & Chiang, S. K. (2017). A pilot study on the Chinese Minnesota Multiphasic Personality Inventory-2 in detecting feigned mental disorders: Simulators classified by using the Structured Interview of Reported Symptoms. *PsyCh Journal, 6*(3), 175–184. https://doi.org/10.1002/pchj.169.

Cheung, F. M. (2004). Use of western and indigenously developed personality tests in Asia. *Applied Psychology. An International Review, 53*(2), 173–191.

Cheung, F. M., Cheung, S. F., Wada, S., & Zhang, J. (2003). Indigenous measures of personality assessment in Asian countries: A review. *Psychological Assessment, 15*(3), 280–289.

Cheung, F. M., Kwong, J. Y. Y., & Zhang, J. (2003). Clinical validation of the Chinese Personality Assessment Inventory. *Psychological Assessment, 15*(1), 89–100.

Cheung, F. M., Leung, K., Fan, R., Song, W. Z., Zhang, J. X., & Zhang, J. P. (1996). Development of the Chinese Personality Assessment Inventory (CPAI). *Journal of Cross-Cultural Psychology, 27*, 181–199. https://doi.org/10.1177/0022022196272003.

Cheung, F.M., Leung, K., Zhang, J.X., Sun, H.F., Gan, Y.Q., Song, W.Z., & Xie, D.(2001). Indigenous Chinese personality constructs: Is the Five Factor Model complete? Journal of Cross-Cultural Psychology, 32, 407–433.

Collazo, A. A. (2005). Translation of the Marlowe-Crown social desirability scale into an equivalent Spanish version. Educational and Psychological Measurement, 65, 780-806.

Constantine, M. G., & Ladany, N. (2000). Self-report multicultural counseling competence scales: Their relatixon to social desirability attitudes and multicultural case conceptualization ability. *Journal of Counseling Psychology, 47*(2), 155–164. https://doi.org/10.1037/0022-0167.47.2.155.

Dana, R. H. (1984). Personality assessment: Practice and teaching for the next decade. *Journal of Personality Assessment, 48*(1), 46–57.

Dana, R. H. (1993). *Multicultural assessment perspectives for professional psychology*. Boston: Allyn & Bacon.

Dana, R. H. (2005). *Multicultural assessment: Principles, application, and examples*. Mahwah, NJ: Erlbaum.

De Marchi, B., & Balboni, G. (2018). Detecting malingering mental illness in forensics: Known-group comparison and simulation design with MMPI-2, SIMS and NIM. *PeerJ, 6*, e5259. https://doi.org/10.7717/peerj.5259.

Edwards, D. W., Morrison, T. L., & Weissman, H. N. (1993). The MMPI and MMPI-2 in an outpatient sample: Comparisons of code types, validity scales, and clinical scales. *Journal of Personality Assessment, 61*(1), 1–18.

Fariña, F., Arce, R., Vilariño, M., & Novo, M. (2014). Assessment of the standard forensic procedure for the evaluation of psychological injury in intimate-partner violence. *The Spanish Journal of Psychology, 17*, E32. https://doi.org/10.1017/sjp.2014.30.

Fernandez, K., Boccaccini, M. T., & Noland, R. M. (2008). Detecting over- and underreporting of psychopathology with the Spanish-language Personality Assessment Inventory: Findings from a simulation study with bilingual speakers. *Psychological Assessment, 20*(2), 189–194.

Fessler, D. M. T. (2010). Cultural congruence between investigators and participants masks the unknown unknowns: Shame research as an example. *Behavioral and Brain Sciences, 33*(2–3), 92.

Fitts, M. S., West, C., Robertson, J., Robertson, K., Roberts, N., Honorato, B., & Clough, A. R. (2015). The NHMRC road map "benchmark" principles: A formal evaluation process is needed to improve their application. *Australian and New Zealand Journal of Public Health, 39*(4), 305–308.

Gierel, M., & Khaliq, S. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. Journal of Educational Measurement, 38, 164-187.

Giger, P., Merten, T., Merckelbach, H., & Oswald, M. (2010). Detection of feigned crime- related amnesia: A multi–method approach. *Journal of Forensic Psychology Practice, 10*, 440–463. https://doi.org/10.1080/15228932.2010.489875.

Hahn, J. (2005). Faking bad and faking good by college students on the Korean MMPI-2. *Journal of Personality Assessment, 85*, 65–73. https://doi.org/10.1207/s15327752jpa8501_06.

Hambleton, R. K. (1996). *Guidelines for adapting educational and psychological tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Hambleton, R. K., & Zenisky, A. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto & F. van de Vijver (Eds.), *Cross-cultural research methods* (pp. 46–74). New York, NY: Cambridge University Press.

He, J., & van de Vijver, F. J. R. (2017). Choosing an adequate design and analysis in cross-cultural personality research. *Current Issues in Personality Psychology, 5*(1), 3–10.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X.

International Test Commission (2005). *ITC guidelines for translating and adapting tests*. Retrieved from www.InTestCom.org. Accessed 30 June 2019

International Test Commission (2016). *ITC guidelines for translating and adapting tests* (2nd ed.). Retrieved from www.InTestCom.org. Accessed 30 June 2019

Kelley, S. E., van Dongen, J. D. M., Donnellan, M. B., Edens, J. F., Eisenbarth, H., Fossati, A., et al. (2018). Examination of the Triarchic Assessment Procedure for Inconsistent Responding in six non-English language samples. *Psychological Assessment, 30*(5), 610–620.

Kiing, J. S. H., Rajgor, D., & Toh, T.-H. (2016). Topical review: Mind your language—Translation matters (a narrative review of translation challenges). *Journal of Pediatric Psychology, 41*(10), 1110–1119.

Kopf, T., Galić, S., & Matešić, K. (2016). The efficiency of MMPI-2 validity scales in detecting malingering of mixed anxiety-depressive disorder. *Alcoholism and Psychiatry Research: Journal on Psychiatric Research and Addictions, 52*(1), 33–50.

Leong, F. T. L., & Pearce, M. (2011). Desiderata: Towards indigenous models of vocational psychology. *International Journal for Educational and Vocational Guidance, 11*(2), 65–77.

Matsumoto, D., & Hwang, H. C. (2013). Assessing cross-cultural competence: A review of available tests. *Journal of Cross-Cultural Psychology, 44*(6), 849–873.

Morey, L. C. (2007). *Personality Assessment Inventory: Professional manual* (2nd ed.). Lutz, FL: Psychological Assessment Resources.

Moultrie, J. K., & Engel, R. R. (2017). Empirical correlates for the Minnesota Multiphasic Personality Inventory-2-Restructured Form in a German inpatient sample. *Psychological Assessment, 29*(10), 1273–1289.

National Health and Medical Research Council. (2018). *Ethical conduct in research with Aboriginal and Torres Strait Islander Peoples and communities: Guidelines for researchers and stakeholders*. Canberra, Australia: Commonwealth of Australia Retrieved from https://www.nhmrc.gov.au/about-us/resources/ethical-conduct-research-aboriginal-and-torres-strait-islander-peoples-and-communities. Accessed 6 July 2019

Nijdam-Jones, A., & Rosenfeld, B. (2017). Cross-cultural feigning assessment: A systematic review of feigning instruments used with linguistically, ethnically, and culturally diverse samples. *Psychological Assessment, 29*(11), 1321–1336.

Nwoye, A. (2015). African psychology and the Africentric paradigm to clinical diagnosis and treatment. *South Africa Journal of Psychology, 45*(3), 305–317.

Öngel, Ü., & Smith, P. B. (1994). Who are we and where are we going? JCCP approaches its 100th issue. *Journal of Cross-Cultural Psychology, 25*(1), 25–53.

Pasick, R. J., Stewart, S. L., Bird, J. A., & D'Onofrio, C. N. (2001). Quality of data in multiethnic health surveys. *Public Health Reports, 116*, 223–243.

Patrick, C. J. (2010). *Operationalizing the triarchic conceptualization of psychopathy: Preliminary description of brief scales for assessment of boldness, meanness, and disinhibition*. Unpublished manual, Department of Psychology, Florida State University, Tallahassee, FL. Retrieved from https://patrickcnslab.psy.fsu.edu/wiki/images/b/b2/TPMmanual.pdf. Accessed 6 July 2019

Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development, 78*, 1255–1264. https://doi.org/10.1111/j.1467-8624.2007.01064.x.

Peters, M. J. V., Jelicic, M., Moritz, S., Hauschildt, M., & Jelinek, L. (2013). Assessing the boundaries of symptom over-reporting using the Structured Inventory of Malingered Symptomatology in a clinical schizophrenia sample: Its relation to symptomatology and neurocognitive dysfunctions. *Journal of Experimental Psychopathology, 4*(1), 64–77.

Rogers, R. (1984). Towards an empirical model of malingering and deception. *Behavioral Sciences & the Law, 2*(1), 93–111.

Rogers, R. (1997). Researching dissimulation. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 398–426). New York, NY: The Guilford Press.

Rogers, R. (2008). Researching response styles. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (3rd ed., pp. 411–434). New York, NY: The Guilford Press.

Rogers, R. (2018a). Detection strategies for malingering and defensiveness. In R. Rogers & S. D. Bender (Eds.), *Clinical assessment of malingering and deception* (4th ed., pp. 18–41). New York, NY: The Guilford Press.

Rogers, R. (2018b). Researching response styles. In R. Rogers & S. D. Bender (Eds.), *Clinical assessment of malingering and deception* (4th ed., pp. 592–614). New York, NY: The Guilford Press.

Rogers, R., & Cruise, K. R. (1998). Assessment of malingering with simulation designs: Threats to external validity. *Law and Human Behavior, 22*(3), 273–285.

Rogers, R., & Gillard, N. D. (2011). Research methods for the assessment of malingering. In B. Rosenfeld & S. D. Penrod (Eds.), *Research methods in forensic psychology* (pp. 174–188). Hoboken, NJ: John Wiley & Sons Inc..

Rogers, R., Gillard, N. D., Wooley, C. N., & Ross, C. A. (2012). The detection of feigned disabilities: The effectiveness of the Personality Assessment Inventory in a traumatized inpatient sample. *Assessment, 19*(1), 77–88.

Rogers, R., Sewell, K. W., & Gillard, N. D. (2010). *Structured interview of reported symptoms, second edition: Professional manual*. Lutz, FL: Psychological Assessment Resources.

Rogers, R., Williams, M. M., Winningham, D. B., & Sharf, A. J. (2018). An examination of PAI clinical descriptors and correlates in an outpatient sample: Tailoring of interpretive statements. *Journal of Psychopathology and Behavioral Assessment, 40*(2), 259–275.

Roth, L. H. (1995). Saleem Shah: His national and international contributions to forensic mental health systems. *Law and Human Behavior, 19*(1), 15–23.

Sánchez, G., Ampudia, A., Jiménez, F., & Amado, B. G. (2017). Contrasting the efficacy of the MMPI-2-RF overreporting scales in the detection of malingering. *The European Journal of Psychology Applied to Legal Context, 9*(2), 51–56. https://doi.org/10.1016/j.ejpal.2017.03.002.

Tellegen, A., & Ben-Porath, Y. S. (2011). *MMPI–2–RF (Minnesota Multiphasic Personality Inventory–2): Technical manual*. Minneapolis, MN: University of Minnesota Press.

Tylicki, J. L., Wygant, D. B., Tarescavage, A. M., Frederick, R. I., Tyner, E. A., Granacher, R. P., & Sellbom, M. (2018). Comparability of Structured Interview of Reported Symptoms (SIRS) and Structured Interview of Reported Symptoms–Second Edition (SIRS-2) classifications with external response bias criteria. *Psychological Assessment, 30*(9), 1144–1159.

University of Minnesota Press, Test Division. (2011). *Guidelines for developing translations*. Minneapolis, MN: University of Minnesota Press, Test Division Retrieved from https://www.upress.umn.edu/test-division/translations-permissions/GUIDELINES. Accessed 6 July 2019

Van de Vijver, F. J. R., & He, J. (2017). Bias and equivalence in cross-cultural personality research. In A. T. Church (Ed.), *The Praeger handbook of personality across cultures: Trait psychology across cultures* (Vol. 1, pp. 251–277). Santa Barbara, CA: Praeger/ABC-CLIO.

van der Zee, K. I., & van Oudenhoven, J. P. (2000). The Multicultural Personality Questionnaire: A multi-dimensional instrument of multicultural effectiveness. *European Journal of Personality, 14*, 291–309. https://doi.org/10.1002/1099-0984(200007/08)14:4<291::AID-PER377>3.0.CO;2-6.

van der Zee, K. I., & van Oudenhoven, J. P. (2001). The Multicultural Personality Questionnaire: Reliability and validity of self- and other ratings of multicultural effectiveness. *Journal of Research in Personality, 35*, 278–288. https://doi.org/10.1006/jrpe.2001.2320.

Van Eeden, R., & Mantsha, R. R. (2007). Theoretical and methodological considerations in the translation of the 16PF into an African language. South African Journal of Psychology, 37, 62-81.

Wooley, C. N., & Rogers, R. (2015). The effectiveness of the personality assessment inventory with feigned PTSD: An initial investigation of Resnick's model of malingering. *Assessment, 22*(4), 449–458.

World Health Organization. (undated). *Process of translation and adaptation of instruments*. Geneva, Switzerland: Department of Mental Health and Substance Abuse Retrieved from http://www.who.int/substance_abuse/research_tools/translation/en/. Accessed 6 July 2019

Zeinoun, P., Daouk-Öyry, L., Choueiri, L., & van de Vijver, F. J. R. (2017). A mixed-methods study of personality conceptions in the Levant: Jordan, Lebanon, Syria, and the West Bank. *Journal of Personality and Social Psychology, 113*(3), 453–465.