Conference Paper

Detecting Anomalies in Assessment Devices for Cross-Cultural Use: Fairness Concerns

Gunadi H. Sulistyo

Universitas Negeri Malang, Indonesia

Abstract

Assessment devices are developed with the undertaking of collecting data that will inform a valid and reliable crux of interest. Such information is mandatory as it will constitute a basis for responsible decision-making. In order to meet the projected function appropriately, any assessment device, let alone those for cross-cultural use, needs to be anomaly-free. While the use of assessment devices across nations has been ubiquitous worldwide, scrutinizing assessment devices for any existing covert defects is unavoidably imperative if fairness in the result interpretation is to be envisioned. This paper reviews briefly several concepts related with nuisances that potentially cause anomaly in assessment devices particularly for cross-cultural use.

Keywords: anomalies, assessment devices, cross-cultural studies

Corresponding Author:
Gunadi H. Sulistyo; email: gunadi.hs@um.ac.id

Received: 1 March 2017
Accepted: 27 March 2017
Published: 12 April 2017

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the LSCAC Conference Committee.

1. Introduction

Assessment devices, be they tests or non-tests, are essentially a set of procedures or instruments that are purposefully designed to reveal the characteristics of subjects or respondents like the factors that deal with cognitive, affective, or psychomotoric aspects [49]. Both tests and non-tests are commonly utilized for classroom or research purposes as a means of collection of data characterizing the subjects' attributes that are manifest or latent in nature.

As a means of data collection, items in assessment devices are by design assigned to measure particular indicators reflecting the existence of underlying characteristics attributable to the subjects. Such items are vital in their role to conceal the true conditions of the subjects' characteristics for the data collected using these items are expected to provide researchers with accurate information about the subjects' attributes. This means that these assessment devices that are naturally comprised of individual items must meet the requirements as good data collection instruments. They must be reliable, valid, practical, and economical [32]. In addition to these, they must also be ecologically sound [33] and fair [19, 38]. By being valid, it means that the devices measure what is designed to be measured. Reliability means that the devices are capable of generating scores that are consistent across time, places, and judges. Practicality refers to ease with which the devices are developed and utilized

and the scores can be interpreted in a simple manner. Being economical means that assessment needs efficiency in terms of spending resources in its development or administration. Ecological soundness means that the information revealed using the devices can be generalized to and reflect the conditions in the real-life settings, cf. [34]. Meanwhile, test fairness means that it is free from any kind of bias.

Meeting these requirements in developing assessment tools is commonly a thought-provoking task. In developing a high quality assessment device, not only does it take rigid procedures that developers need to exert but also extraordinary expertise that not everyone has. Despite all the efforts exercised and spent for developing the instrument, the result may not be as satisfactorily as expected. If creating an instrument is such a challenge, adapting an existing instrument is a possible way out. However, adapting an existing instrument needs to be performed appropriately by considering several aspects if adaptation becomes the choice let alone adapting the one once used in a population that differs culturally from the intended population of interest.

In another context, however, it is interesting to differentiate two types of cross-cultural studies based on orientation [44, 46] as structure and level orientation studies. The former deals with the studies seeking evidence whether an instrument measures the same construct in different cultures; meanwhile, the latter is concerned with cross-cultural differences that are evidenced in average scores. Thus, there are studies that are deliberately designed and carried out to examine similarities or differences in conceptualization of particular areas across cultural populations speaking their own languages, for instance religiosity among Asians and Europeans. Such studies use instruments to collect data from people with their own languages. Adapting instruments is therefore unavoidable. In order for the scores obtainable from different populations to be comparable, equivalence in the adapted instruments needs to be maintained. Otherwise, bias in measurement will result, which can have serious impacts in a number of vital ways: breaking equal chances for getting jobs or having education, or drawing invalid conclusions in research, cf. [4].

This paper deals with the need to recognize anomalies in measurement across cultures with an emphasis on equivalence and bias. In such a context it is frequently inevitable to adapt an existing data collection instrument that has different cultural context of use and is intended for people speaking a different language of different studies. To that end, several subtopics that follow are presented: concept of equivalence, types of equivalence, concept and types of bias, detecting bias, and concluding remarks.

2. Concept of Equivalence

In general the term equivalence is closely related with the idea of comparability. In the context of instrument adaptation, equivalence, according to [46], refers to the condition in which test scores are comparable. These scores are generated by administering two measures: an original data-collection instrument and the other set of instrument that has been adapted from the original instrument and is administered in a different cultural group. The more the scores are comparable or share similarities to the scores generated by means of the original instrument from which the adapted instrument is made, the more equivalent the original instrument and the adapted instrument will be. [15] terms the condition resulting from the adapted instrument as 'fair'. For instance, upon adapting AMTB (Attitude/Motivation Test Battery) developed by [7] for use in Indonesian context, a researcher obtained scores that are comparable to the scores generated from the original AMTB. The administration of the adapted instrument is said to be equivalent to that of the original instrument. Consequently, it can be stated that the measurement using the adapted instrument is fair, in that it does not put the Indonesian respondents at the disadvantage because of being measured by using the adapted instrument. The administration of items in the AMTB does not yield biased scores.

Based on the discussion of equivalence above, it can be interpreted roughly that "... bias, in this view, is synonymous to nonequivalence; conversely, equivalence refers to the absence of bias" [46], p. 145. However, [48], p. 493, contend that "... bias and equivalence are not inherent characteristics of an instrument, but arise in the application of an instrument in at least two cultural groups and the ensuing comparison of scores, patterns or item values." "Bias refers to nuisance factors that jeopardize the validity of instruments applied in different cultures. Equivalence refers to the level of comparability of scores across cultures" [46], p. 145. In a way, nevertheless, the presence of bias may be considered as evidence of presence of factors resulting in inequivalence in the features of instruments as well as administration of measures of interest.

3. Types of Equivalence

A thoroughgoing discussion on the concept of equivalence can be found in the work of [16] who identifies fifty detailed terms for equivalence, which he later classifies into two main categories: interpretative equivalence and procedural equivalence. By interpretative equivalence he means whether the constructs that are measured across cultures share similarities and differences in their interpretation or meaning. Procedural equivalence is concerned with the nature of equivalence in measures and procedures employed cross-culturally in studies across cultures.

Taking another perspective, [44] as cited in ([47], p. 122) distinguish equivalence into three levels: construct equivalence, measurement unit equivalence, and full score equivalence. For the present discussion and practicality, this classification of equivalence by [44] is considered. What follows is a brief account of each.

Construct equivalence, being the lowest in level, is also synonymously termed as measurement invariance [43] or measurement equivalence [6]. It deals with correspondence in the use of the theorizing, or conceptualization of same latent traits - what to measure - as a basis to develop the instrument utilized across cultural populations, cf. [26, 39], [45]. As construct equivalence reflects the idea of "....[whether] measurement operations yield measures of the same attribute." ([10], p. 117), which essentially deals with construct validity of a measure [3, 25], construct equivalence is necessarily an important element that characterizes a data-collection device [6].

A modest example in English learning context would be the term 'reading comprehension' for the lower secondary level of education in Indonesian context and that of Japan as perceived in the concept of reading competences in their corresponding syllabus. Conceptually, reading comprehension in both syllabi differs markedly. In Indonesian context, for a lower secondary level student to be considered able to read, he/she must be conceptually able to demonstrate the mastery about nine microskills of reading: topic, main idea, specific information, detailed information, reference, inference, purpose, tone, and vocabulary skills [40]. Although some competences may overlap, they differ from those of Japanese context. This can be extended to the case in which any other standardized test, for instance TOEFL, is used to test students' learning achievement.

Measurement unit equivalence is also called metric equivalence ([9], p. 8). It takes place when measures have equivalence in the measurement unit (of interval or ratio scales), but a difference in the onset. To illustrate, to measure the length of ropes, we can use either a unit of meter or that of inch. For instance, in one group there are plastic ropes and metal ropes; in another group there are also plastic ropes and metal ropes. If the ropes in the first group are measured using the meter unit, the scores of the plastic ropes and metal ropes in the first group are comparable. Similarly, if the ropes in the second group are measured using the inch unit, the scores of the plastic ropes and metal ropes in the second group are also comparable. The scores of the plastic ropes and the metal ropes in each group are comparable as the measurement unit used is the same. However, the scores of the measurement across the first and the second group are not as directly comparable as the scores in both groups do not have metric equivalence. "With metric equivalence, scores can be compared within cultural groups..., and mean patterns and correlations across cultural groups, but scores cannot be compared directly across groups" ([9], p. 8).

Another simple example might be the case in which two ethnic groups within which there are different educational levels are tested using the TOEFL and the IELTS. The

ethnic group A, for instance, is measured using the TOEFL; the other ethnic group, group B for instance, using the IELTS. The scores of the different educational levels within each group are directly comparable. However, the scores of both groups are not directly comparable as the TOEFL and the IELTS employ different scoring systems.

Full score equivalence is also termed as scalar equivalence ([9], p. 8), which constitutes the highest level in equivalence. This type of equivalence is characterized by the use of the same measurement unit and the same onset. Therefore, the results of measurement, if full score equivalence is satisfactorily met, are directly comparable as the scores obtained are free from bias. In ([9], p. 9) argue that construct equivalence can be achieved if the construct is bias-free. Also, construct equivalence is not affected by method and item bias. However, bias in both methods and items affects both measurement unit equivalence and full score equivalence. Therefore, for measurement unit equivalence and full score equivalence to be achieved conceptually, methods and items to collect data should be methodologically designed to be free from bias.

4. Concept of Bias and its Types

Bias is evident when there is variance between the scores produced by different subjects of the same ability in responding to items measuring the same underlying construct. In ([47]: 120) states that bias materializes "... when score differences on the indicators of a particular construct do not correspond to differences in the underlying trait or ability." ([9], p. 8) echo [44, 47] on bias as coming from three sources: construct bias, method bias, and item bias. What follows is a brief account of each topic.

Construct bias happens when the construct under consideration across culture is defined partially or when the criteria characterizing conducts in each culture related to the construct under study are not all appropriately addressed or considered in an instrument for cross-cultural measurement ([9], p. 5). This means that the construct under measurement includes unidentical elements or partial factors of each culture that is considered in the instrument of interest. An example of this would be the term 'language learning anxiety'. Is this term conceptualized well to include identically defined areas across cultures? A common practice in the data collection in the context of research about English learning is the use of a set of questionnaires developed by [11] known as the Foreign Language Classroom Anxiety Scale (FLCAS). The instrument is originally used in studies in the English speaking context, for instance, a study by [31]. However, this instrument is also commonly used to assess students' level of learning anxiety across cultures, for instance in the study by [1, 20, 50]. If the instrument only addresses unidentical dimensions or their corresponding factors, the instrument by definition is said to be biased in its construct. A study by [50] investigated the stability of the general foreign language classroom anxiety construct across English, French, Japanese and Russian among Chinese undergraduate foreign language learners as

measured using FLCAS. This study reveals that the levels of general foreign language anxiety are likely contingent on which native language and which foreign language are learned and the general foreign language anxiety varies across languages according to the specific target language to be learnt. The findings obviously suggest the need to use instruments that are fair and culture-sensitive in research across cultures. Therefore, care should be taken when using assessment devices that originate from an instrument once used to assess features of a particular cultural group.

A concept similar to construct bias is the so-called 'construct underrepresentation' - the term coined by Embretson ([46], p. 145). Construct underrepresentation takes place when concept mapping of a construct is not performed comprehensively so that some dimensions or variables of the construct under investigation are not well represented by their corresponding indicators. In other words, only few indicators from which items are derived are employed to measure a construct which essentially has a broad coverage. An example of this case is the incomprehensiveness in representing the construct of reading comprehension in reading comprehension tests according to graduate standard competences of a syllabus. In spite of including all micro skills of reading in the test as stipulated in the syllabus, the test includes only few indicators to assess students' reading comprehension.

The next kind of bias is that of methods, or method bias. [9, 44, 45] outline three sources of method bias: sample bias, instrument bias, and administration bias. What follows is a brief account of each of these types of bias. Characterizing features of a particular cultural group necessitates the data that representatively reflect the group. This means that all members with their aspects related to the group need to be carefully put into account when these members are to be involved as the source of data. The researchers need to identify carefully first the characteristics of the data source prior to selecting the individuals that will become the basis for information collection. For instance, data sources may be spread across vast geographical areas with different administrative managements. In such a case, the researcher needs to ponder data sources located in the city, district, or the sub-district. On the other hand, a researcher may need also to come up with the idea of data richness or abundance in individuals if data sources are meant to be echoed from them. In the second case the data source is determined arbitrarily on the basis of such data richness or abundance. A study that aims at explaining how successful learners of English reach their achievement in taking the TOEFL, for instance, will consider merely those who have demonstrated high achievers in the TOEFL scores as the data source. Failure to think over the characteristics of data sources can lead to sample bias. Therefore, sampling procedures, be they of probability or non-probability, need to be carefully designed and exercised as well as accurately carried out.

Instrument bias deals with the disturbance related with the performance of the respondents due to the use of particular instruments. There are sources of instrument

bias: stimulus familiarity, response procedures [9, 47], and response styles [46]. Familiarity with a stimulus deals with the situation in which the materials in the instrument advantages one group of respondents but it disadvantages another while these two groups share equal traits to be assessed. An example of this bias results from the use of pictures in a speaking test or an interview as stimuli to elicit test takers' spoken competence. If the pictures contain a situation not existing in one of the different groups' culture, thus causing a group's unfamiliarity with the situation, the pictures most likely prevent one group of the test takers from producing relevant competences meant to be revealed by the material. A reading comprehension passage, for instance, may contain materials of highly specific cultural materials. Or, the comprehension questions following in the passage utilize meta-language known only by certain groups to assess the test takers' understanding of the passage. These situations are examples of bias due to familiarity with a stimulus.

Response procedures constitute another kind of instrument bias. An example of this response procedure is the one as shown in a study by [22] who are interested in examining the effects on performance of computer familiarity and attitudes towards CB IELTS. The study reveals that the test takers felt that those who were more advanced in computer skills accomplish the CB IELTS tasks better than those with only basic skills in computer operations. Formats of a test familiar to respondents affect their performance on the test. Respondents who frequently encounter a particular type of instrument formats will find it relatively more facilitating in accomplishing the task than those who do not. Another piece of research by [30] reveals that the test takers' performances on a multiple choice section with which they are more familiar performed better than those who did the formats involving true-false and gap-fill in the blank formats to which they are not accustomed.

Response styles make up another source of instrument bias. Cronbach, as cited by ([9], p. 6), refers to response styles as "a systematic tendency to use certain categories of the answering scale on some basis other than the target construct". This includes acquiescence response styles (ARS) which are '...tendency to agree rather than disagree to propositions in general' (Lentz, as quoted by [9], p. 6) willingly without objections or protests, cf. [12, 14, 24] contend that acquiescence response styles stem from levels of intelligence. Less educated individuals are inclined to agree than disagree to statements in a questionnaire. Personality and beliefs also constitute another source of bias due to acquiescence response styles. Those who are cautious have the tendency to agree with statements in questionnaires, even more likely to accept rather than question them. As a result, they tend to have prejudice and intolerance to these kinds of statements [18].

The last type of method bias is administration bias. In [46] lists several sources of administration bias. The first is the conditions within which the instrument of interest is administered. These conditions may refer to the physical or the social ones. A listening

comprehension test administered in different testing environments with a different quality of the audio devices used to generate the auditory input, for instance, constitutes a kind of administration bias. Participants assessed in a room with a low quality of aural recording will be deprived as a result. Similarly, crowded seating arrangements in a testing room will lead to another kind of administration bias.

Administration bias may also potentially take place in a condition where instructions for the test takers to undertake the test are not clear; or, it also can happen in the situations in which the procedures or the protocols for administrators to collect the data are confusing, or are partially followed. Another potential circumstance for administration bias to take place is when a test administration involves a number of administrators with different levels of expertise. For instance, a number of interviewers are engaged in a selection interview. When the interviewers involved come up with unequal levels of ability in interviewing, some candidates may be assessed in an unfortunate situation that results in bias.

Another source of administration bias is the so-called tester/interviewer/observer effects. A very common phenomenon in such a kind of bias is what is known as halo effects. Halo effects happen when an attribute or characteristic of a person is inappropriately transferred to another condition as a basis to make an overall evaluation of that person. As an example, someone's native-like use of a language may be used by an interviewer to judge the person's other good attributes, for instance, the person's good IQ, and vice versa. An EFL learner who speaks with an English accent perfectly and fluently may be believed to have smart intelligence by an interviewer. A quick judgment caused by halo effects in such a situation results in administration bias. Finally, administration bias may also take place when there are constraints that happen during an interview between an interviewer and an interviewee. These problems take place in a range of situations from language gaps between the interviewer and the interviewee to prejudice – 'an unfair feeling of dislike for a person or group because of race, sex, religion, etc.' (`http://www.merriam-webster.com/dictionary/prejudice`) - and stereotypes – 'a generalization, usually exaggerated or oversimplified and often offensive, that is used to describe or distinguish a group' of people (`http://www.dictionary.com/browse/stereotype`) - about the interviewee raised by the interviewer. In [17] conducted a study to investigate whether there are interactions between gender stereotypes for jobs, applicant gender, and the communication styles used by male and female applicants during an interview. The study demonstrated that male respondents got a penalty on ratings of overall impression and hire ability for communicating in stereotypically gender-inappropriate manners; meanwhile female respondents got penalized on ratings of sociability and likeability for communicating in a stereotypically gender-inappropriate fashion. On the other hand, in a simulated study involving an interview [36] reveal that stereotype threat disadvantages females

more than males. These studies imply that during an interview, communication bias potentially occurs.

The final type of bias is the so-called item bias. There are different phrasings of item bias. In [35], p. 4, for instance defines it '...as a multifaceted component of observed test scores that, like the usual measurement error, causes an observed score to be different from the "true" score, but, unlike the usual measurement error, is associated with membership in a particular group and, for members of that group, has an expected value other than zero.' Item bias is also defined as '...differential probability of answering an item correctly when having the same ability but belonging to different subgroups.' ([23], p. 389). Meanwhile, [21], p. 3, refer to item bias as "....differences in item-level responses that occur when items function differently for certain groups of respondents". These definitions imply that item bias is problematic phenomena in measurement processes as an item with bias in it potentially produces a score that does not actually reflect a test taker's or a respondent's true ability or skill. An item is said to be biased when the performance of test takers or respondents in responding to an item is differential while these respondents have the same level of knowledge or skills being measured in the item of interest. However, not all differences in the mean score of the respondents' score indicate the presence of bias caused by the item as random errors in measurement may cause this. Item bias takes place systematically by disadvantaging one group or an individual over another one whose underlying trait to be measured is equal irrespective of their religious affiliation, gender, socio-economic status, and ethnicity.

There are sources of causes of item bias. In ([46]: 146) notes the sources as "poor translation and/or ambiguous items, nuisance factors, and cultural specifics" contained in the item of interest. Several studies, for instance Fitriani's study (Fitriani, 2015), adapt instruments once originally used in a study involving a group of English speakers. Assumed that the use of this original instrument may cause misunderstanding on the part of the respondents whose mother tongue is not English, the original instrument is then translated into another language version of the target respondents. Thus, this translation is expected to facilitate respondents or test takers in responding to the items of interest. However, if the translation quality is poor, the items in the newly adapted instrument may endanger respondents' or test takers' responses as these may become the potential sources of bias in its items. An ambiguous item can also lead to item bias. Consider the following item:

Children like chewing candies. So does my sister, especially....

1. red candies, green candies, and blue candies
2. red, green, and blue candies

The item obviously poses an ambiguity in its alternatives. Do the candies come from different groups of different colors? Or, are the candies three colored?

The other source of item bias is 'nuisance factors'. These factors are irrelevant factors not intended to be assessed in items, but accidently contained in an item. As a result, such items "...may invoke additional traits or abilities ..." [4] that are in theory to interfere with the testing of a particular competence of test takers. A test of grammatical items is meant to assess the test takers' knowledge of grammar, not other abilities. However, if the item is phrased using vocabulary items of low frequency that may hinder the test takers' grammatical competences being assessed, such an item is biased. Another example would be a case of a test of reading comprehension. A reading comprehension item by design assesses the test takers' understanding of the content of the passage. In other words, the test takers' response to the item should be not only relevant but also 'hooked' on the passage content. The test takers should be dependent on the passage content in responding to the item, not on their general knowledge about the content. Conversely, a test intended to assess test takers' general knowledge involving a reading passage —such a test is commonly called an integrated test— should not test reading comprehension skills of the test takers.

The last source of item bias is 'cultural specifics.' These include for instance "incidental differences in connotative meaning and/or appropriateness of the item content" ([46]: 146). Studies of culture may be focused on aspects or constructs of a particular culture with its cultural specificity. Or, they seek similarities or differences of aspects or constructs across cultures. The former is commonly referred to as the study of 'emic' aspects; the latter 'etic' aspects of cultures [2] – ideas coming from the study of human speech sounds as phonemics – the study of speech sounds of a particular language – and phonetics – the study of speech sounds of language in general. This by concept implies that not all aspects of a culture may not then exist in another culture; or aspects of a culture may also be found in another culture. In assessment context it is argued that when materials that are of emic nature related to the content domain [13] are tested across cultures through items, item bias may occur. However, at the level of item, [29] provides types of assessment bias: offensiveness and unfair penalization and its possible sources categorized as differential groups: racial/ethnic bias, gender bias, and socioeconomic bias. A hypothetical example of specificity related to ethnic bias would be a writing test prompt used nationally in Indonesian context that requires test takers to write a descriptive text about Borobudur Temple. Although Borobudur Temple is known as a national heritage among Indonesians and it is taught in history classes to students, this does not automatically mean that all students know it in details. If the prompt is to elicit students' writing competences in describing its detailed features, biased scores may result. Students may know language features used for describing. They may also know the generic structure of descriptive texts. However, when it comes to describing the temple details such as the materials, not all students may be able to do so satisfactorily due to lack of knowledge in cultural specificity of the temple with which they are suppused to be familiar.

5. Detecting Bias

Bias is essentially a critical annoyance in measurement as it potentially informs fallacious competences, or skills and knowledge that respondents try to demonstrate by means of administering tests or any other data-collecting assessment devices. If truth collected from respondents as the data source is projected, any form of bias should be avoided at all points. Therefore, such tests or any other data-collecting assessment devices must be bias-free so that they can function fairly across respondents or test takers regardless their gender, ethnicity, religion affiliation, and other cultural attributes. Detecting bias then constitutes a critical and imperative phase in developing data-collecting assessment devices.

Detection of bias constitutes a part of instrument development stage. Essentially bias can be detected through two strategies: conceptual and empirical. The conceptual strategy is approached using expert validation/review; meanwhile, the empirical one is performed by involving a tryout to collect data. Expert validation may be carried out using three ways: think aloud, independent panel, focused group discussion [1], p. 207, in which a series of questions that [8] develop may be utilized. The form provides a useful review form with which to analyze whether an item is biased or not.

Empirical validation can be performed using both classical test theory (CTT) and item response theory (IRT), usually involving statistical analysis [5, 37]. Under the tradition of CTT, several techniques for detecting bias are commonly used. These are among others the Transformed Item Difficulty (TID) method or the delta approach, the Mantel–Haenszel (MH) method, and the Logistic Regression (LR) method. Meanwhile, in the IRT, several methods are used. These are the Lord's Chi-square method, and the Raju's Area method.

6. Conclusions and Suggestions

Anomaly in assessment across cultures is existent and becomes a thoughtful concern in test development. In ([13], p. 4653) outlines areas of concerns in this area. The outline is useful in that it may serve as a basis for assessing cultural elements. However, when cross-cultural use of assessment devices in need is in place, bias and equivalence should be considered carefully and situated accurately. With this in mind, fairness in measurement is assured. In this perspective, cross-cultural studies may be concerned with anyone of the following approach: the indigenous, the cultural, and the cross-cultural approaches. Within any approach, equivalence should be exercised and optimized; while bias should be kept as minimum as possible. Based on the ideas proposed by [46, 47], equivalence may embrace these: construct equivalence, measurement unit equivalence, and full score equivalence, whereas bias may deal with construct bias, method bias, and item bias. Technology to detect bias is available there.

References

[1] F. A. Bachtiar, E. Cooper, G. H. Sulistyo, and K. Kamei, "Student assessment based on affective factors in English learning using fuzzy inference," in *International Journal of Affective Engineering*, vol. 15, pp. 101–108, 2015.

[2] J. W. Berry, "Imposed Etics-Emics-Derived Etics: The Operationalization of a Compelling Idea," International Journal of Psychology, vol. 24, no. 1, pp. 721–735, 1989.

[3] J. D. Brown, "What is construct validity?" JALT Testing & Evaluation SIG Newsletter, vol. 4, no. 2, p. 12, 2000.

[4] A. Brown, "Measurement invariance and differential item functioning. A short course in applied psychometrics," Peterhouse College, vol. 10, p. 12, 2012.

[5] G. Camilli and A. L. Shepard, Methods for identifying biased test items, Sage, London, UK, 1994.

[6] E. Davidov, B. Meuleman, J. Cieciuch, P. Schmidt, and J. Billiet, "Measurement equivalence in cross-national research," Annual Review of Sociology, vol. 40, pp. 55–75, 2014.

[7] R. C. Gardner, "Attitude/motivation test battery: International AMTB research project," 2004, http://publish.uwo.ca/%7Egardner/docs/englishamtb.pdf.

[8] R. Hambleton and J. Rodgers, "Item bias review. Practical Assessment," Research & Evaluation, vol. 4, no. 6, p. 1, 1995, http://pareonline.net/getvn.asp.

[9] J. He and de. F. Van Vijver, "Bias and equivalence in cross-cultural research," Article, vol. 8, p. 19, 2012.

[10] J. L. Horn and J. J. McArdle, "A practical and theoretical guide to measurement invariance in aging research," Experimental Aging Research, vol. 18, no. 3, pp. 117–144, 1992.

[11] E. K. Horwitz, M. B. Horwitz, and A. J. Cope, "Foreign language classroom anxiety," Modern Language Journal, vol. 70, no. 2, pp. 125–132, 1986.

[12] C. H. Hui and H. C. Triandis, "The instability of response sets," Public Opinion Quarterly, vol. 49, no. 2, pp. 253–260, 1985.

[13] G. Iskifoglu, "Approaches to study culture, equivalency and bias: Triadic notions in developing research instruments for cross-cultural comparative studies," in *Proceedings of INTED2014 Conference 10th-12th*, pp. 4653–4658, Valencia , Spain, 2014.

[14] D. N. Jackson and S. Messick, "Response styles on the MMPI: Comparison of clinical and normal samples," Journal of Abnormal and Social Psychology, vol. 65, no. 5, pp. 285–299, 1962.

[15] A. R. Jensen, "An examination of culture bias in the wonderlic personnel test," Intelligence, vol. 1, no. 1, pp. 51–64, 1977.

[16] T. P. Johnson, "Approaches to equivalence in cross-cultural and cross-national survey research," ZUMA-Nachrichten Spezial, January 1998.

[17] J. L. Juodvalkis, B. A. Grefe, M. Hogue, D. J. Svyantek, and W. DeLamarter, "The effects of job stereotype, applicant gender, and communication style on ratings in screening interviews," in *and communication style on ratings in screening interviews. The International Journal of Organizational Analysis*, vol. 11, pp. 67–84, 2003.

[18] E. S. Knowles and K. T. Nathan, "Acquiescent Responding in Self-Reports: Cognitive Style or Social Concern?" Journal of Research in Personality, vol. 31, no. 2, pp. 293–301, 1997.

[19] A. J. Kunnan, "Fairness and justice for all. In Antony," in Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, John. Kunnan and John. Antony Kunnan, Eds., p. 14, University, Orlando, Florida, 2000.

[20] A. K. Liauw, Language anxiety in speaking among students in Khims English course (Unpublished Masters Thesis [Master, thesis], Nommensen University, Pematang Siantar, North Sumatera, 2012.

[21] L. Marotta, L. Tramonte, and J. Douglas Willms, "Equivalence of testing instruments in Canada: Studying item bias in a cross-cultural assessment for preschoolers," Canadian Journal of Education, vol. 38, no. 3, 2015.

[22] L. Maycock and T. Green, "The effects on performance of computer familiarity and attitudes towards CB IELTS," Research Notes, vol. 20, p. 3, 2005.

[23] C. D. Mccauley and J. Mendoza, "A Simulation Study of Item Bias Using a Two-Parameter Item Response Model," Applied Psychological Measurement, vol. 9, no. 4, pp. 389–400, 1985.

[24] G. Meisenberg and A. Williams, "Are acquiescent and extreme response styles related to low intelligence and education?" Personality and Individual Differences, vol. 44, no. 7, pp. 1539–1550, 2008.

[25] S. Messick, (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd Ed.) (pp. 13-103). New York: Macmillan.

[26] R. E. Millsap, "Statistical approaches to measurement invariance," Statistical Approaches to Measurement Invariance, pp. 1–355, 2012.

[27] com. Merriam-Webster. Merriam-Webster and n. d. Web, 2016.

[28] S. Osterlind, Test Item Bias, vol. , SAGE Publications, Inc., 2455 Teller Road, Newbury Park California 91320 United States of America, 1983.

[29] J. W. Popham, Assessment Bias: How to Banish It, Person, Boston, 2012.

[30] O. Rezaei, H. Barati, and M. Youhanaee, "The effect of content familiarity & test format on iranian efl test takers' performance on test of reading comprehension," International Journal of Applied Linguistics and English Literature, vol. 1, no. 4, pp. 1–14, 2012.

[31] M. Rodriguez and O. Abreu, "The stability of general foreign language classroom anxiety across English and French," Modern Language Journal, vol. 87, no. 3, pp. 365–374, 2003.

[32] J. Salvia and E. J. Ysseldyke, Assessment, Houghton Mifllin Company, Boston, 8th edition, 2001.

[33] R. J. Sbordone, "Ecological validity: Some critical issues for neuropsychologists," in Ecological validity of neuropsychological testing, R. J, C. J. Sbordone, and Long., Eds., pp. 15–42, St. Lucie Press, Boca Raton, 1998.

[34] M. A. Schmuckler, "What is Ecological Validity? A Dimensional Analysis," Infancy, vol. 2, no. 4, pp. 419–436, 2001.

[35] J. Scheuneman, "Exploration of Causes of Bias in Test Items," ETS Research Report Series, vol. 1985, no. 2, pp. i–73, 1985.

[36] A. Shantz and P. G. Latham, "The effect of stereotype threat on the interview performance of women," Advancing Women in Leadership, vol. 32, no. 1, p. 29, 2012.

[37] L. A. Shepard, G. Camilli, and M. D. Williams, "Validity of approximation techniques for detecting item bias," Journal of Educational Measurement Vol, vol. 22, no. 2, pp. 77–105, 1985.

[38] E. Shohamy, "Fairness in language testing," in Fairness and Validation in Language Assessment: Selected papers from the 19th Language Testing Research Colloquium, A. J and Kunnan., Eds., pp. 15–19, University, Orlando, Florida, 2000.

[39] J.-B. E. M. Steenkamp and H. Baumgartner, "Assessing measurement invariance in cross-national consumer research," Journal of Consumer Research, vol. 25, no. 1, pp. 78–90, 1998.

[40] G. H. Sulistyo, Reading for meaning, Pustaka Kaiswaran, Malang, 2011.

[41] G. H. Sulistyo, "EFL learning assessment at schools., Malang," Bintang Sejahtera, 2015.

[42] C. Barker, The SAGE Dictionary of Cultural Studies, vol. , SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, 2004.

[43] R. J. Vandenberg and C. E. Lance, "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research," Organizational Research Methods, vol. 3, no. 1, pp. 4–69, 2000.

[44] K. Leung and F. J. R. Van de Vijver, Methods and data analysis for cross-cultural research, Sage, Newbury Park, CA, 1997.

[45] F. J. R. Van De Vijver and Y. H. Poortinga, "Towards an integrated analysis of bias in cross-cultural assessment," European Journal of Psychological Assessment, vol. 13, no. 1, pp. 29–37, 1997.

[46] F. J. R. Van de Vijver, "Bias and equivalence: Cross-cultural perspectives. In," in Cross-cultural survey methods, J. A. Harkness F and P. Ph. Mohler, Eds., pp. 143–155, Wiley, New York, NY, 2003.

[47] F. J. R. Van de Vijver a and N. K. Tanzer, "Bias and equivalence in cross-cultural assessment: an overview. Revue europeenne de psychologie appliquee: an overview. Revue europeenne de psychologie appliquee 54," pp. 119–135, 2004.

[48] F. J. R. Van de Vijver and R. Fischer, "Improving methodological robustness in cross-cultural organizational research," R. S. Bhagat and R. M. Steers, Eds., pp. 491–517, Cambridge University Press, New York, 2009.

[49] M. Williams and R. Burden, Psychology for language teachers, Cambridge University Press, Cambridge, 1997.

[50] G. Yan, "The stability of general foreign language classroom anxiety across languages among Chinese undergraduate foreign language learners," Journal of Asia TEFL, vol. 7, no. 2, pp. 69–89, 2010.