INTERNATIONAL TEST COMMISSION

ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations

Version 4.2

Please reference this document as:
International Test Commission. (2018). *ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations.* [www.InTestCom.org]

ACKNOWLEDGEMENTS

The ITC thanks René Lawless (USA) and María Elena Oliveri (Canada) who drafted the guidelines and served as committee chairs for this project.

The ITC also thanks members of the project committee for their contributions to the draft. These included Avi Allalouf (Israel), Sydell Carlton (United States), Thomas Eckes (Germany), Paula Elosua (Spain), Molly Faulkner-Bond (USA), Ronald Hambleton (USA), Dragoş Iliescu (Romania), Stephen Sireci (USA), Fons van de Vijver (Netherlands), Alina von Davier (USA), and April Zenisky (USA).

The ITC thanks its members of the ITC Council as well as Cathy Wendler and Robert Mislevy, who provided useful feedback on earlier versions of the guidelines.

SUMMARY AND SCOPE

These guidelines describe considerations relevant to the assessment of test takers in or across countries or regions that are linguistically or culturally diverse. The guidelines were developed by a committee of experts to help inform test developers, psychometricians, test users, and test administrators about fairness issues in support of the fair and valid assessment of linguistically or culturally diverse populations. They are meant to apply to most, if not all, aspects of the development, administration, scoring, and use of assessments; and are intended to supplement other existing professional standards or guidelines for testing and assessment. That is, these guidelines focus on the types of adaptations and considerations to use when developing, reviewing, and interpreting items and test scores from tests administered to culturally and linguistically or culturally diverse populations. Other guidelines such as the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) may also be relevant to testing linguistically and culturally diverse populations.

These guidelines are designed to inform test developers, psychometricians, and test users of the considerations that should be made to help ensure test fairness and score comparability to support meaningful inferences in culturally and linguistically diverse contexts. They augment existing ITC guidelines and other professional guidelines (or standards), referenced at the end of this document. Although they primarily apply to large-scale assessments administered in education, their general principles may also apply in other settings such as licensure, certification, and tests of skill mastery (such as those for a driver's license). In small scale or one-to-one (clinical) assessments, there may be challenges to the implementation of these guidelines. Oakland (2016) provides specific recommendations for best practice as they apply to the administration of these kinds of assessments to individuals who are immigrants and/or second-language learners.

There is much sensitivity surrounding the terminology used in reference to the diverse languages used in a country or region. Henceforth, these guidelines try to limit the use of terms as minority and majority language, or native and foreign language, to refer to the various languages of an assessment for a country or region. Some of the many factors that influence linguistic diversity are illustrated in the Introduction. In these guidelines, references to linguistic groups may refer to cultural and/or historical second language (L2) speaker groups as well, based upon the context of the standard. This terminology has been chosen for the sake of efficiency.

A glossary of uncommon or technical terminology can be found at the end of this document. **Underlined terms throughout this document are linked to the glossary. To access these definitions, press *Ctrl* and the hyperlink.**

CONTENTS

INTRODUCTION

An important element in addressing fairness issues in assessment entails consideration of how to accommodate the linguistic needs of test takers within linguistically diverse countries or regions. Such contexts may be the result of migration (for economic, social, political, or religious reasons); interests for the maintenance and revitalization of other languages; past colonization; and/or other conditions that allow people to move from one region/country to the other. With linguistically diverse populations, various considerations are required, particularly if test takers' home language(s) are different from the language used in the school, community, or test. One of the difficulties that may occur is the identification of a test taker's dominant language. There may be additional considerations in countries where there is more than one official language.

These guidelines are intended to inform test developers, psychometricians, test users, and test administrators about fairness issues in support of the fair and valid assessment of linguistically or culturally diverse groups. These guidelines are meant to apply to most, if not all, aspects of the development, administration, scoring, interpretation and use of assessments and are intended to supplement other existing professional standards or guidelines. Because assessments are used to inform diverse decisions (some of which are high stakes), the guidelines cover considerations for the breadth of the lifecycle of an assessment; that is, from its conceptualization to its implementation and interpretation of scores.

Factors Affecting the Fair Assessment of Linguistically or Culturally Diverse Populations

Central to the development or adaptation of fair and valid assessments for linguistically or culturally diverse populations is the consideration of the contextual factors that influence test taker's response processes. Individual test takers' proficiency or mastery of different languages may be due to differences in their language acquisition processes. For groups of test takers, other differences that may be related to *societal dimensions* also merit consideration. The language used by an institution may differ as it relates to the official or unofficial recognition of the test taker's language. These various factors have implications on the availability of resources and curricular materials, the degree of teacher training in that language, and importance given to revitalization of the language. There might also be differences in the degree to which languages are *codified*. For instance, there are languages that have only an oral tradition; that is, they have not been codified, the codification is recent, or they are in the process of codification.

Legal status of the diverse languages within countries

Some countries give official status to more than one language, which means that those languages can be used within public institutions. Even if a language is not widely spoken, it can have a legal status. For instance, New Zealand has three official languages: English, Maori,

and New Zealand Sign Language. In Belgium there are three official languages: French, Flemish (Dutch), and German. Spanish is the national language of the country of Spain but it has other five official autonomous regional languages: Galician, Basque, Aranese, Valencian, and Catalan. The official status of a language affects the ability to secure resources, financial support, and the creation/adaptation of new educational materials, extending the language to new domains and into the training of teachers.

Language of instruction

An important moderator of test performance is the language of instruction used in schools. When the language of instruction is not the same as a test taker's home language(s), it is important to decide which should be the testing language in order to accurately assess his/her performance on the construct and preserve the validity of score-based inferences. This helps to disentangle test takers' knowledge of an assessed construct from their linguistic proficiency level.

Choosing the appropriate language in which to assess test takers in such cases is complex. On one hand, there is the question of whether an assessment is available in the test taker's dominant or home language(s). On the other hand, there is the question of whether the test takers have enough knowledge of the testing language to enable valid score-based inferences about their knowledge of the assessed content or construct without the possibility of confounding it with their knowledge of the language of the test.

A further complication exists in making the decision of which language should be chosen for such test takers. Is it the test taker's native language, the language of the domain that is to be assessed, or the language, in the case of educational assessments, in which they were taught the subject matter? Questions also arise regarding who should make that decision and what criteria they should use to make those decisions. These decisions may involve asking questions about what is feasible and practical, and whether there are resources available to assist these decisions.

A very complex situation in choosing a testing language arises across linguistically or culturally diverse countries. To illustrate, South Africa has 11 official languages, not all of which are treated similarly. Some might receive greater levels of support in education and public life as compared to others. Since English is the common language of the country, it may make it the preferred testing language. However, the considerable variation in educational quality across schools in the country leads to such variation in English proficiency among test takers that it cannot be easily used as the language of assessment without prior verification of test takers' mastery of English.

Institutional support

The availability of resources within a country depends on, among other things, the legal status of a language, a country's affluence, and the prestige of the different languages that coexist. In some contexts, tests from other countries may be used. There are, however, limitations related to the appropriateness and relevance of the use of the original normed scores; maintenance of the construct definition or curricular relevance across groups; or the comparability of scores for the multiple test-taker populations. Consideration of these issues is important because they may affect the accuracy of score-based interpretations. Similar considerations may apply to the use of instruments developed for use in a language (e.g., English) for one country, and then using it in other countries that speak the "same" language.

Codification

Many languages are codified with an alphabet or written code and others are not. For example, in Morocco there are languages (e.g., Arabic, French) that have official status and are codified; but there are others (e.g., varieties of Berber) that exist only in verbal form. Differences in linguistic codification or status elevate the difficulty in creating suitable test adaptations for the multiple populations.

Social Status and Prestige of Languages

Linguistic differences may also exist in relation to the *social status*, or *prestige* of languages in contact with the mainstream language, which (combined with political or cultural dominance in certain cases) may reflect diverse attitudes towards the various languages.

Languages in multilingual societies can differ in their social status and prestige. The *prestige* of a language(s) describes the level of respect accorded to a language(s) or dialect as compared to other languages or dialects in a speech community. Having an account of the sociolinguistic context of a language can help an educational agency choose the language of an assessment as well as interpret any possible score differences among linguistic groups.

Naturally, there is not a common and unique context that clearly defines the characteristics of linguistically or culturally diverse populations, since they vary across and within countries. Thus, simple "one size fits all" recommendations will have limited value.

Multiple language uses

A further complication with some languages (e.g., Arabic and Chinese) arises out of whether the language is spoken (spoken Arabic) or whether it is standardized (written Arabic). In the case of Arabic, there are regional variations in word usage. In the case of Chinese, multiple languages are spoken yet they all use the same character-based written system. Consequently, common words in a spoken language may not always be identical to the written form.

Other differences may occur at the regional level. For example, in Canada, there might be test takers who speak French or English depending on whether they live in an Anglophone or Francophone community. Some test takers might be born in Canada, while others might be new immigrants. Alternatively, test takers may have been in transit for a number of years with interrupted schooling in the language of instruction. Due to the high levels of immigration in Canada, regional differences might also occur across French-speakers, for example, because they come from multiple French-speaking countries, some of which might have different variations of French including Creole.

The availability of resources for adapting tests

The availability of resources depends, among other things, on a country's affluence, the official status of the languages of incoming immigrants (within a region or country), and how well organized the immigrant populations are. In some cases, tests from other countries may be used. However, the use of such tests (and their normative scores) may be problematic as the instruments are typically not intended for use outside of the country in which they were developed, may measure different content or constructs, or have inappropriate data with which to compare performance. Similar challenges may arise in the use of English instruments across English-speaking countries, which may differ in their use of particular terminology or phrases. It may not be possible to correct or adjust for these differences based upon available resources.

THE GUIDELINES

Guideline 1: Test Development and Adaptation

Adaptation of Existing Tests for Linguistically or Culturally Diverse Populations[1]

1.1. Test developers/publishers should consider the linguistic differences between the source and target languages and cultures (grammatical, syntactical, semantic, lexical, etc.) when adapting tests or other instruments for test takers of the targeted language, in order to make the forms as psychometrically comparable as possible. Special attention should be given when the languages belong to different linguistic families.

1.1.1. Individuals from different linguistic groups should be involved in the design of the items and the test to be adapted as they are best suited to identify any translation hurdles that may occur and make suggestions on how to circumvent those hurdles.

1.1.2. Cultural aspects should be considered when translating the components of the test (items, scales, rubrics etc.) and efforts should be made to adapt the test from source to target not only linguistically, but also culturally.

1.2. When necessary, adapt the wording of the item language (from the source language) for the targeted language test takers to be assessed, provided that it does not change the construct.

1.2.1. The test in the targeted language should be similar in length to the source test and each item should contain the same number of option choices as in the source items.

1.2.2. The items in the targeted language should be of the same register as the source items, the same level of difficulty, and not include connotations that are absent from the original text.

1.2.3. Because word-for-word translations of items may not make sense in the targeted language, the translations should convey synonymous ideas related to the construct without altering the difficulty of the item.

[1] For a high-level overview of test translation and adaptation principles, see the ITC Guidelines for Translating and Adapting Tests (2nd edition), https://www.intestcom.org.

1.3. When a test is developed for several cultures and/or languages, or is adapted for a target culture and/or language, consider that the format of the items, stimuli, scoring rubrics, and test instructions are equally familiar for all target populations.

1.3.1. Sampling procedures should be as similar as possible for the different language and/or culture test forms to prevent sample bias during the analysis of equivalence.

1.3.2. Differential familiarity with the stimulus material, differential response styles, or differential social desirability for the different-language and/or different-culture populations may lead to instrument bias. These differences should be investigated during the analysis of equivalence.

1.4. When a test is adapted, ensure that the placement of elements on the page such as the pictures and page numbers do not interfere with the readability of the text. Review all figures in adapted items for suitability for all the linguistic groups.

1.4.1. The layout of the test in the targeted language should be as close as possible to the source. For example, test takers in the targeted language should not be at a disadvantage because they need to turn the page or scroll the document while the entire text appears on the same page for the test takers in the source language. In addition, for right-to-left languages, make sure that the images are mirrored (or not) depending on the country's usage.

1.5. All adapted tests should be evaluated for accuracy by reviewers (fluent not only in the source and target languages, but also in the source and target cultures) to ensure fidelity to the construct and proper translation. Any adaptations made should be documented and provided to the test user.

Item Format Familiarity and Design of Test Instructions

1.6. Design test instructions to maximize clarity (e.g., use simple and clear language).

1.6.1. Present test instructions using various modalities (e.g., oral and written form); and, where possible, give instructions in the test takers' dominant language, as long as the test is not assessing language proficiency.

1.7. Test developers/publishers should provide evidence (such as editorial or fairness reviews) that the language used in the test instructions, rubrics, and test items is clear for the users of the test as well as the test takers.

1.8. Do not assume that the L2 test takers have previous experience with given task types or item types. Rather, evaluate the familiarity of item formats to ascertain they are appropriate for all test takers, regardless of their linguistic group. Item

formats that are appropriate for all the populations should be preferred over formats that need to vary from one group to another.

Item Development and Review

1.9. When selecting topics for items, avoid topics that may be considered offensive, derogatory, or exclusionary, or may cause an emotional reaction from members of any one of the linguistically or culturally diverse populations, as it may create construct-irrelevant bias.

1.10. Develop test items and reading passages that contain accessible vocabulary for all linguistic and cultural groups. Language should be used that is free of any regional and sensitive vocabulary. Also, avoid words with multiple meanings or other unnecessarily complex words that are not part of the assessed construct.

1.10.1. When items are written for several linguistic groups, native speakers from each language group should be consulted to ensure that problematic terminology is avoided, such as those that are regional, or sensitive to particular populations.

1.11. Where possible, avoid the use of ambiguous language in the source language version of the test such as the use of truncated stems in the prompts, as it may be difficult to develop adaptations of such terms in various languages.

1.12. When not part of the assessed construct, develop items that use a simple sentence structure. Several shorter sentences are often preferred to a single more complex sentence.

1.13. Where possible, develop items that use contexts that depict common scenarios for all linguistic groups and populations.

1.14. Unless it is part of the assessed construct, avoid reference to historic context and names that might be well known to some cultures but not to others.

1.15. Keep the language or reading demands of the items on a test to the minimum necessary to assess the construct of interest.

1.16. Avoid the use of construct-irrelevant product names, entertainers, geography, government, holidays, measurement systems, and currency, among others that might be relevant or more familiar to only some cultural/linguistic groups.

1.17. When adapting items into the targeted language versions, pay special attention to identifying and avoiding wording that may have different meanings for different linguistic groups.

1.18. **Reviews by experts in each language group should be provided to help ensure that the items cover the intended construct or domain for all the linguistic populations. The experts who review or select items from an item pool should be knowledgeable of the culture of the different linguistic groups and be fluent in the language of the items they review. Ideally, these experts should be native to the targeted language and culture.**

1.19. **Test items, rating scales, and test materials should be reviewed for any elements (e.g., historical events, situations, pictures, colors) for which members of the diverse linguistic populations may be sensitive or unacquainted. The use of linguistic/cultural assessment experts is suggested for conducting these reviews during the initial item development stage.**

1.20. **When feasible, background demographic questions that are to be developed for the test should sensibly ask about the test takers' language background with enough detail to enable meaningful analyses at the group level.**

Item Tryout

1.21. **If possible, conduct item try-outs or cognitive interviews with test takers from all linguistic groups to ascertain the appropriateness of the items for each language population and determine if the test takers from each linguistic group are interacting with the items in the desired way.**

1.22. **If an item on an assessment is not language based, such as an equation or a picture stimulus, and no adaption is required, provide empirical evidence regarding the comparability between the first language (L1) and second language (L2) learner groups in the population.**

1.23. **For any items where it is found that some linguistic group test takers are not using the intended response processes, evaluate the item (including any visual prompts, instructions, rubrics, etc.) to identify any edits that may make the item clearer and more functional for them.**

1.24. **If there is a sufficient amount of data for all linguistic groups, conduct statistical analyses to ensure that the items do not function differently among diverse linguistic populations.**

Guideline 2: Validity, Reliability, and Fairness

Validity

2.1. When an adapted version of a test is used as an accommodation for any test takers, the degree to which it is comparable to the original test version should be evaluated.

2.1.1. Validity studies should be conducted to ensure that an adapted version of a test measures the intended construct(s), based upon its intended purpose.

2.1.2. An equating/linking model between the versions should be considered, including anchor items at the design level and common persons during a pretest trial or administration.

2.2. Ensure that relationships between test scores and other variables are comparable between all linguistic and cultural groups.

2.2.1. Validity evidence based on relations of test scores to other variables can provide important evidence regarding how well test scores meet their intended purpose. When testing linguistically or culturally diverse populations, the degree to which these relationships hold up for subgroups of test takers, such as those defined by linguistic or cultural diversity, should be studied. Any differences found in correlation or prediction may require further investigation and/or documentation to investigate possible unintended consequences.

2.3. If different versions of the items are developed for test takers from different linguistic groups, these changes should be documented, and the invariance of their psychometric characteristics should be included in the documentation, including any impact of the changes on score interpretations.

2.4. If score interpretations are allowed to vary across linguistic groups (e.g., separate norm tables for groups defined by country or language), a rationale should be provided for permitting the variations, and document the impact of the variations over test score interpretations and uses.

2.4.1. When full invariance cannot be established between different language forms of the test, partial invariance is an acceptable compromise. Partial invariance establishes invariance not for all items, but for subsets of items. If analyses are developed on such subtests, invariance may be upheld, but the content coverage of the test may be impacted and therefore significance of the scores may differ. If this is the case, documentation on such variations should be provided.

2.5. Evaluate the invariance of the internal factor structure of the assessment across the L1 and L2 populations.

2.6. If it is found that test takers who take the test in different languages are passing the test at different rates relative to one another, support this finding with other forms of empirical evidence, to show that such differential rates are not due to bias in the test's construction or scoring.

2.6.1. If scores of different linguistic populations are found not to be comparable, provide evidence that such differences will not cause adverse impact on test scores. If it does cause adverse impact, present strong evidence that the intended use is served, without unintended negative consequences.

2.7. The test and testing mode should consider the range of abilities of all test takers appropriately, including the different linguistic populations of test takers.

2.7.1. To assess the range of abilities of all test takers appropriately, the variance of the ability distribution between the different linguistic populations of test takers should be considered.

2.7.2. When choosing the testing mode (i.e., computer adaptive testing, multistage adaptive testing, modular, or linear testing) the range of abilities of the L2 test takers, and their familiarity with testing modes, should be considered.

2.7.3. If the score differences in the different linguistic populations are large, consider selecting an adaptive test (rather than a linear test) to increase measurement precision for all test takers in an efficient manner.

Construct Relevance

2.8. The relevance of the construct measured for both the L1 and L2 test takers should be documented. Such documentation should include judgmental arguments made from a sociocultural perspective and arguments based on empirical evidence, (i.e., evidence that the validity of score interpretations is equivalent between all linguistic and cultural groups).

2.9. When evidence for construct relevance for linguistically or culturally diverse test takers is based on expert reviews, the characteristics of the sample of participants or expert judges should be documented.

2.10. For assessing a test taker's (L2) language proficiency, use a separate language assessment. Where possible, administer this test annually (using different equated test forms) as test takers' level or proficiency may change from one year to the next.

2.10.1. Determine test taker proficiency for the most appropriate language for the test unless the measured construct is language proficiency.

Reliability

2.11. Ensure that the scores from assessments meet acceptable reliability criteria for every linguistic population, both in an absolute sense, and relative to the original population.

2.11.1. Where appropriate, conduct reliability analysis to help ensure an acceptable minimum reliability standard for all linguistic groups. Examples of such analyzes include coefficient alpha, test-retest reliability, test information functions, classification/decision consistency, standard errors of measurement, and conditional standard errors around performance standard cut scores.

Fairness

2.12. Conduct fairness reviews for all items and test elements (including prompts, instructions, images, rubrics, etc.) with a focus on linguistically or culturally diverse groups of test takers.

2.12.1. Include representatives from all linguistic and cultural groups in panels of language experts used for fairness reviews.

2.12.2. To the extent possible, all test materials should be judged to be free of:

- Offensive or overly generalized portrayals of the different linguistic populations;
- Images or references that are likely to be unfamiliar to test takers from the different linguistic groups and are not directly relevant to the construct of the assessment;
- Images or phrases that are offensive in other cultures or religions;
- Language, imagery, or content that is likely to be unnecessarily advantageous or disadvantageous for test takers from different linguistic populations.

2.13. Include test takers from all linguistic and cultural groups in the norming group (if scores will be norm-referenced) or experts from all linguistic and cultural groups on the standard setting panel (if scores will be criterion-referenced) to ensure all

linguistic and cultural groups are represented in the determination of performance standards for the assessment.

2.13.1. Materials provided to the test user should include descriptions of the standard setting panel and the normative data, if available. This should include information about the demographics of the norming population and the date when the assessment was normed.

Score Comparability

2.14. In the case of adapted tests, conduct score comparability studies to examine the extent to which test scores are invariant for both versions of the test.

2.14.1. If validity evidence indicates scores from adapted and original tests are comparable, they should be treated in the same way as all other scores.

2.14.2. If validity evidence indicates scores across original and adapted tests are non-comparable, then (a) the steps taken to ensure comparability should be reviewed, and (b) the procedures for adapting tests should be revised. Further, experts should examine the non-comparable items to determine whether they can be further adapted to establish comparability with the original test.

2.14.3. If validity evidence indicates scores across original and adapted tests are non-comparable, test users should be informed of the non-comparability using readily available documentation.

2.15. Provide a clear rationale and supporting evidence to demonstrate that scores between different language test forms are comparable. This includes scoring rubrics, well-defined rater training, and use of statistical models (such item response theory) to evaluate comparability.

2.15.1. To support score comparability across the linguistic and cultural groups, provide evidence of measurement invariance of equated scores, including Differential Item Functioning (DIF) analyses, when there is an adequate sample size.[2]

[2] For more information about various procedures that can be used to analyze DIF, see the "Confirmation Guidelines" section in ITC's Guidelines for Translating and Adapting Tests (Second Edition), https://www.intestcom.org.

2.15.2. Where relevant, provide detailed technical information about the method selected for equating/linking the scores on two language versions of the test.

2.16. If the scores of a test are to be norm-referenced, the various linguistic groups should be represented in the norming group in the language version of that test. The norming should be based on the population of the region and linguistic varieties of where the test will be administered.

2.17. If significant and systematic differences in score comparability across language groups are found, an investigation (such as a linguistic/cultural analysis) should be conducted to help ensure the differences will not result in score discrepancies that disadvantage any linguistic group of test takers.

Examining Sources of Differential Item Functioning

2.18. Conduct differential item functioning (DIF) analyses for each item as appropriate for the item type and assessed construct to ensure that test takers from the different linguistic groups are not impacted differentially on items relative to test takers from the reference group.

2.18.1. If sample sizes are adequately large, identify salient subpopulations within each linguistic group and consider running DIF analyses for each subpopulation.

2.19. Evaluate item content for sources of DIF to investigate possible sources of construct-irrelevant variance.

2.19.1. Where possible, have linguistic/cultural assessment experts conduct reviews of items that are flagged for DIF.

2.19.2. If the source of DIF is found to be irrelevant to the assessed construct, consider revising or removing the item.

Guideline 3: Scoring Essays and Other Constructed-Response Items

Developing, Designing, and Adapting Scoring Rubrics

3.1. The design of the scoring rubric should reflect the linguistic diversity of the targeted population as well as the purpose of the assessment.

3.1.1. Scoring rubrics should be developed so that they do not unfairly penalize non-native speakers of a language or test takers who interpret items differently within the context of a specific cultural background. For example, if test takers are answering science, short constructed-response items, their fluency in the language of the assessment should not affect the scoring unless it interferes with comprehensibility.

3.1.2. Scoring notes should be created for raters so there is no penalty for language spelling patterns, and limited language proficiency or cultural difference is not mistaken for limited content knowledge.

3.2. Different scoring designs may impact test takers from different linguistic groups in various ways.

3.2.1. When there may be differences in the specific aspects or facets of the assessed construct between the linguistic groups, use scoring that distinguishes between these different aspects (analytic/trait scoring).

3.2.2. When using holistic scoring for an overall judgment of interrelated skills, ensure that the overall performance score is comparable between the linguistic groups.

3.3. In consideration of the potential heterogeneity of responses between different linguistic populations, pretest the design of the rating scale with a representative sample of test takers from the overall population, including test takers from all the linguistic groups.

3.3.1. Conduct analyses to gain insight into the effectiveness and utility of each of the rating scale categories (for example, the levels of endorsement). Where possible, conduct these analyses with all linguistic populations to see whether differences arise between groups with respect to the use of the rating scale categories.

3.3.2. If the difference in rubric use for some linguistic groups is judged to have an impact on their outcomes, it may be necessary or appropriate to rescore items after the revision of the rubric and/or rater recalibration. The purpose of the rescoring is not to artificially increase or reduce the score of a specific

linguistic group but to examine whether the differences are due to construct-relevant sources.

Scoring Materials

3.4. Include topic notes and benchmark examples for all linguistic groups to describe potentially different stylistic writing patterns, highlighting those that might lead to placing the essay at a lower point in the scale for construct-irrelevant reasons.

3.5. All scoring should be done anonymously. Background information about test takers should not appear on the material to be scored, including their name; country of origin; age; sex; language background; or ethnic or cultural membership.

3.6. When scores are assigned by multiple raters, ensure that the other raters' scores are not visible or otherwise identifiable on the form.

Rater Selection and Training

3.7. Clearly define the qualifications and characteristics of new raters and select them based upon these qualifications. Ideally, raters should have previous experience with scoring a wide range of performances by test takers from the different linguistic groups.

3.7.1. When using task-specific or trait (analytic) scoring, use raters who are familiar with and understand the different linguistic populations, when possible.

3.8. As a group, raters should represent a broad spectrum of demographic, regional, content, and professional backgrounds and, where possible, include members of the linguistic populations, who can resolve possible score discrepancies that may (dis)advantage test takers from L2 groups due to construct-irrelevant sources. These raters can also help recalibrate other raters to help ensure fairness for all populations.

3.9. To ensure an efficient, well-organized scoring process, highly trained raters experienced with responses from all the linguistic groups should conduct and oversee the scoring, as scoring or table leaders. These leaders are responsible for monitoring the scoring behaviors of the other raters and for ensuring fidelity to the scoring rubric.

3.10. Provide raters with a sufficiently large and varied sample of practice responses that are atypical of the targeted population, including benchmark responses from all the linguistic populations.

Benchmarks and Rangefinders

3.11. The raters representing different linguistic groups should be given precisely defined criteria to score the responses, analyze the context of the items, and resolve discrepancies among raters. The criteria may be used in training and recalibrating other raters. Use pre-scored responses (benchmarks) to exemplify each rating scale category or level descriptor of the scoring rubric, including responses from test takers from each linguistic group. Use these benchmark responses to evaluate raters' alignment with the scoring rubric to certify raters are successfully calibrated with the benchmarks.

3.11.1. Use rangefinders to help raters consensually define the category intervals and illustrate responses in regions of the rating scale that are important to give raters a better understanding of the distinctions between score points, particularly when test takers from different linguistic groups are concerned.

3.12. Examine whether there are any discrepancies occurring in raters' scores assigned to test takers' responses from each linguistic group and discuss these discrepancies in relation to whether they arise out of construct-relevant or construct-irrelevant sources.

Scoring Plan

3.13. When a group of trained and certified raters (each with experience in scoring tasks for test takers from at least one linguistic group) is available, a decision should be made regarding the number of raters to be employed in operational scoring sessions. Irrespective of the number of raters per response, the raters should submit independent scores in order to avoid unwanted effects such as when two or more raters negotiate the scores or imitate each other.

3.14. In the scoring plan, consider the following constraints: time schedule, budget, the importance of the assessment outcomes for test takers (e.g., high- vs. low-stakes decisions), the level of reliability required, and the scoring design (including the way raters are assigned to test takers, tasks, and performances). For example, when an assessment consists of multiple tasks, reliability is higher when different raters score a test taker's performances than when the same rater scores each test taker's performances.

3.15. Devise a scoring plan that is both cost- and time-efficient and still allows the scoring leader to compare all raters, test takers, and tasks within the same frame of reference.

3.16. Strive to achieve a scoring plan that links raters, test takers, criteria, and tasks, while being mindful of linguistic diversity. A network of links is a prerequisite

for taking into account, for example, differences in the level of severity or leniency each individual rater exhibits when assigning scores to test takers.

3.17. When scoring responses, pair speakers of different languages to compensate for any potential bias associated with each rater's perspective or point-of-view.

Inter-Rater Reliability and Agreement

3.18. Use indices of inter-rater reliability and agreement to quantify the extent to which raters disagree with one another in order to provide evidence on the overall success of rater training procedures for all linguistic groups.

3.19. Compute at least two different rater agreement statistics: one consensus index, indicating the degree to which raters assign the same or similar scores to the same responses (e.g., the percentage of exact or adjacent agreements) and one consistency index, indicating the degree to which raters consistently rank-order test takers' responses (e.g., the Pearson correlation).

3.20. In the case of high-stakes decisions, reliability requirements are particularly strict; so use at least two independent raters to make judgments on final assessment outcomes.

3.21. Compare indices of inter-rater reliability and agreement between operational scores (if two or more raters provide scores for the same set of performances) or compute the indices between operational scores and scores provided by expert staff or table leaders.

Rater Monitoring and Evaluation

3.22. Regularly monitor rater scores to maintain consistent and accurate scoring throughout scoring sessions, particularly in high-stakes assessments under conditions of heterogeneous populations.

3.22.1. Scoring or table leaders should employ read-behind or read-ahead quality check procedures, if available, particularly in online scoring programs.

3.22.2. Utilize read-ahead procedures to identify agreements and disagreements between operational and expert scores.

3.22.3. Include results of read-behind or read-ahead procedures in individual raters' summary quality reports.

Rater Recalibration and Retraining

3.23. In the case of significant deviations from quality expectations or standards set by the scoring leader, retrain or recalibrate raters manifesting unacceptably low scoring quality using new sets of practice responses, benchmarks, and rangefinders.

3.24. Re-assign recalibrated raters to operational scoring sessions only if quality checks indicate a sufficiently high rate of agreement with other operational raters and with scoring leaders.

Rater Statistics and Feedback

3.25. Over the course of operational scoring sessions, regularly collect and analyze information provided by raters (i.e., holistic scores, analytic subscale scores, total scores, category usage, etc.) to derive rater statistics.

3.26. Use statistics such as rater means, rater standard deviations, score or scale category frequencies, and agreement with other operational raters or with scoring or table leaders to give feedback to individual raters on their scoring behavior, for example, in cases of overly lenient or harsh scoring.

3.27. Calculate rater agreement statistics and inter-rater reliability coefficients to provide evidence on the extent to which each rater deviates from, or is aligned with, the other raters.

Approach to Disagreement Resolution

3.28. When two or more raters provide discrepant scores for the same set of test-taker responses, a method of resolving rater disagreements needs to be specified and used to report a single score for each test taker.

3.28.1. Resolution methods include averaging the two original scores (rater mean), incorporating the scoring of a third rater during adjudication (parity method), or replacing both original scores by the score of an expert adjudicator (expert method). The choice of a particular method will depend upon time and budget considerations, as well as on the availability of highly experienced, expert raters.

3.28.2. The raters who represent different linguistic groups can resolve possible score discrepancies that may (dis)advantage test takers from L2 groups due to construct-irrelevant sources.

Measurement Results and Analysis of Rater Effects

3.29. If available, build on expertise in psychometric modeling of observed scores in order to closely monitor, analyze, and evaluate scoring behaviors in assessments of all the linguistic groups.

3.29.1. Study various sources of measurement error (e.g., dialect differences, different task formats, or different language versions of the same test), for example use Generalizability theory, to estimate the magnitude of each source, and to provide a strategy for improving the reliability of the test or assessment.

3.29.2. Estimate an ability measure for each test taker as free as possible of the particularities of the test or assessment situation, for example, to compensate for rater effects such as differences in rater severity or leniency for any linguistic group. For this purpose, use a method such as Many-Facet Rasch Measurement (MFRM).

3.30. Examine potential differences that may arise between the different linguistic populations possibly arising due to differences in (human or automated) scoring modes.

3.30.1. Ensure that test takers from different linguistic groups are not penalized for differences in writing styles that might differ from the reference population for reasons not central to the assessed construct.

Guideline 4: Test Administration Procedures and Instructions for Accommodations

Test Administration

4.1. The test administration manual should specify all aspects of the test administration that require scrutiny for new linguistic or cultural populations. It should be written in the language in which the test will be administered.

4.1.1. Test proctors should read verbatim any scripts provided for test administration in the language of the test or in the language of each of the test-taker groups.

4.2. When possible, administer the test in the test taker's most proficient language, unless language proficiency is part of the assessment.

4.3. If any linguistic test taker groups are to be tested on alternate dates, each language group should have a similar number of testing dates offered in a year.

4.4. Describe linguistic adaptations and their rationale in detail in the test administration manual, as recommended by the test's publisher.

4.5. The test administrator should follow best practices related to test administrations[3] with all groups. In addition, the test administrator is responsible for the following activities relevant to testing all linguistic and cultural groups prior to the testing session:

- Those aspects of the physical environment that influence the administration of a test or instrument should be made as similar as possible across populations of interest, as indicated in the test administration manual. If it appears that there may be a situation unaccounted for in the test administration manual, the test administrator should either contact the test user or test developer for advice, or make an effort to organize the testing to minimize any disruptions that may occur for test takers receiving accommodations.
- Helping to ensure that the test administration manual specifies all aspects of the test administration that may require scrutiny in terms of accommodations that may need to be made for members of new linguistic or cultural groups.
- Ensuring that test proctors administering the test received proper training and are sensitive to the needs of test takers from all linguistic groups.

[3] For more information, see the ITC Guidelines on Test Use: https://www.intestcom.org

- Advising test takers from all linguistic groups of the linguistic or dialectic groups for which the test is considered appropriate.

4.6. The test manual should contain clear, explicit, and easy to understand instructions for the test proctors to reduce sources of error and make clear the test takers' rights and responsibilities.

4.6.1. Test proctors should explicitly follow the procedures contained in the manual for the test administration.

4.6.2. Where possible, test proctors should give the test instructions in the primary language of the test takers to minimize the influence of unwanted sources of variation across populations, even where the test content is designed to provide evidence of knowledge or skills in a non-primary language. Test proctors should read verbatim any scripts, instructions, or examples provided for test administration in the language of the test or in the language of each of the test taker groups.

4.6.3. To maintain standardized conditions and avoid sources of variation, it is not suitable that the proctor defines personal criteria for instructions or explanations, including hints.

4.7. Test proctors should explain test takers' rights and responsibilities. Test proctors should be sensitive to a number of factors that can impact the validity of the inferences drawn from the scores for all populations, related to the stimulus materials, administration procedures, and response modes.

Test Accommodations

4.8. All test accommodations should be developed and documented to allow for the valid measurement of the targeted construct for members of all linguistic and cultural groups.

4.9. Prior to any test administrations, determine the kinds of allowable accommodations that can be made for test takers from any cultural or linguistic group in the administration of the assessment that will not alter the construct measured. Procedures regarding the ways in which the accommodations will be implemented should be established prior to test administration.

4.9.1. Accommodations may include, but are not limited to, the use of (electronic) dictionaries, the translation of lexicons with difficult words into the different languages, additional time, glosses, providing test instructions in the test taker's home language(s), alternate test forms administered in the test taker's home language(s), or the use of interpreters.

4.9.2. If interpreters are used, they should be fluent in both the test taker's native language as well as in the language of the test. Ideally, the interpreter should be experienced in translating between the two languages and understand the unintended consequences that could be caused by poor translations.

4.10. When tests are to be administered to individual test takers, consider accommodations for test takers on an individual basis in advance, since it may be the case that each person may possess a different degree of acculturation or competence with the test language.

4.11. During all phases of the testing process, treat all test takers who will be receiving accommodations fairly and similarly to other test takers.

Guideline 5: Score Interpretation and Reporting

Score Interpretation and Score Reporting

5.1. **Develop score reports and accompanying interpretive materials for the various linguistic and cultural populations taking the test, particularly in the case where results have important consequences for individual test takers.**

5.2. **A review of all existing score reports and interpretive materials should be conducted to inform the planning and development of any new materials**

5.3. **Score report developers should be aware of differences in language status among the intended recipients of the reporting materials (i.e., test takers, their parents, or the local authorities). Reporting efforts should consider how score reports and related materials could be communicated to maximize understanding and usefulness to all score report recipients. For instance, information could be transmitted in various forms, such as verbally, visually, as well as in written form to reach various audiences including users with limited literacy levels or groups without written languages.**

5.4. **Use focus groups and other data collection methodologies (think-alouds, interviews, observations, etc.) to identify confusing elements of score reports (e.g., technical terms, complex sentences) that are potentially problematic to any linguistic group that are present in the testing population.**

5.5. **Gather evidence for the interpretability and use of score report materials for all linguistic populations when such materials are used, and use these data to inform score report revisions and future reporting practices.**

Score Report Design and Contents

5.6. **In score report documents, use language that is appropriate to the score report recipients to communicate the test purposes and appropriate test uses. Pay special attention to those linguistic or cultural populations who may be less familiar with test score scales, reports, and interpretive guides.**

5.7. **In the score reports, include interpretive guides or information for the reference population as well as for the different linguistic populations, particularly for the most prevalent language populations. Parallel forms of the score reports and interpretative guides across the different languages should be the goal. If possible, translate the score reports into each different language, and describe technical terms related to scoring to ensure that the translators have a working understanding of any technical terminology that is related to scoring.**

5.8. Design score reports so that key results are visually prominent and understandable across the different language populations.

5.9. Consider test purposes and uses to inform choices about numerical, graphical, and text-based results to include in the score reports. The choices made should consider the relevance of the different types of results for the intended recipients, their linguistic proficiency, and technical considerations about the levels of detail supported by the data.

5.9.1. Where graphic displays of results are used, clearly label numerical scores, score scales, and other display elements using simple and direct language that is appropriate for all linguistic populations.

5.10. Carefully consider group-based comparisons between test takers based on demographics (geographical, linguistic, race/ethnicity, etc.) in light of the test content to avoid possible potential misinterpretations of observed patterns. In reporting materials prepared for all linguistic and cultural populations, communicate results in clear, non-judgmental language and provide explanations of the results to avoid possible misinterpretations of the data.

5.11. When reporting results, including subscores, test developers have a responsibility to communicate the technical caveats (for example, higher measurement error than the overall score) related to the subscores. They should also communicate what interpretations of the subscores are appropriate for all linguistic populations.

Delivery of, and Access to Score Reports and Interpretive Materials

5.12. The translation/adaptation of score reports and interpretive materials into minority languages present in the testing population should be carried out by qualified translators, incorporating best practices for ensuring score meaning and appropriate interpretations.

5.13. Provide score recipients of score reports and ancillary interpretive materials in their preferred language or provide them with clear information about how to obtain these reports.

5.14. When developing ancillary interpretive materials for each represented language, the test developers/testing publisher should ensure that the actual delivery mechanism enables access to those materials for all users from the different linguistic and cultural populations.

Test Use

5.15. Provide guidance about score meaning and uses in clear and straightforward language that reflects the linguistic levels of intended users.

5.16. Select a test based upon its suitability for the test purpose while taking into account the test and the background characteristics of the targeted population, including all linguistic groups.

5.16.1. If the test is a norm-referenced test, an examination of the norming sample should be conducted to ensure that it is representative of the targeted population, including all linguistic groups.

5.17. The use of a test is under the responsibility of the developer and the user. Both must provide sufficient evidence about the purpose and the use of the test according to its design and construct.

5.17.1. It is the responsibility of the developer of an assessment to communicate the intended uses of the test and interpretations of its scores clearly, with evidence to support the claims made about the construct for the diverse populations taking the test.

5.17.1.1. If resulting score interpretations are inconsistent with the construct the assessment is intended to measure, the test user is responsible for strongly cautioning stakeholders about the lack of evidence to support the claims that might be made with respect to linguistic or cultural groups.

5.17.1.2. The test user has the right to cancel or invalidate scores when an assessment is utilized for an unapproved use of the assessment or its scores.

5.17.2. It is the responsibility of the test user to document the rationale for the selection of a particular assessment, including evidence to support the claims about the construct that can be made about that assessment.

5.17.3. If a deviation from the specified purpose of the test is desired, the test user must provide a rationale and evidence to support the new use of the assessment, including an explanation regarding the interpretation of the resulting scores for each linguistic group. A validity study should be conducted to ensure that the assessment supports the intended use.

5.17.3.1. When evidence exists of inappropriateness of the new use of the assessment, any of the stakeholders (including test takers, test administrators, or score users) should report the test misuse to the test user

that oversees the decisions that are made, based on scores from that assessment.

5.18. Wherever possible, if the population of test takers is culturally or linguistically diverse, the developers and publishers of the assessment should provide clear information regarding the appropriate and inappropriate uses of the test and the interpretation of scores.

5.19. The test developer should provide appropriate technical guidance regarding the types of adaptations that were made to the assessment; instructions on score interpretation; for whom the adapted assessment is intended; and how the validity of the inferences made from the scores may be.

5.20. The test developer and the user of the test should strive to have a clear understanding of scores, their validity, and the impact those scores will have on test takers from any of the linguistic and cultural groups.

5.20.1. The user of the test is responsible for monitoring and ensuring that potential misinterpretations or inappropriate uses of the test do not occur under their authority.

5.20.2. For test results that are publically released, the user of the tests should ensure that appropriate explanatory materials are provided to the public to avoid misinterpretation, particularly if there are differences in the results of the test takers from the different linguistic groups.

5.21. If any major adaptations are requested by the test user and are made to the test in terms of its format, language, or mode of administration, the user of the test should have a sound rationale. In addition, the user of the test should conduct validity and reliability studies on the modified assessment.

Guideline 6: Quality Control to Achieve Comparability and Fairness in the Scoring of Tests

6.1. To help ensure standardized testing conditions, consider creating a checklist to help ensure that all scoring processes use the appropriate rubrics for the different linguistic and cultural groups.

6.2. For rater-scored assessments, consider evaluating raw scores to determine whether there are interactions between the raters and responses from some linguistic groups. If significant and systematic differences in scores are found, an investigation should be conducted to ensure that the score differences will not have an unintended negative consequence on test takers from those linguistic groups.

6.3. The quality control of scale scores should be carried out prior to the final reporting of scores for both the L1 and L2 language populations.

6.3.1. The technical manual of the test should include an explanation about the methods used to define the scale and its equivalence to the raw scores, and, if needed, the form the user may utilize to adapt the scale for a different context.

6.4. Where possible, examine the scores for each linguistic population separately and then compared to each other as well as to the reference population.

6.4.1. Where possible, compare the expected and observed scores for each linguistic group against each other to look for trends (based on scores from previous test administrations and the current administration).

6.4.1.1. Check for score precision, reliability, and speededness across populations and test administrations.

6.4.1.2. Differential speededness between the L1 and L2 test takers may signal that the scores may not be comparable.

6.4.2. Evaluate score changes for test takers from any linguistic group as well as score changes for the total group.

6.5. Document all statistics and analysis for future quality control studies and checks.

6.6. Check the means and standard deviations of score changes for test takers who have taken the assessment more than once.

6.6.1. As the assessment continues to be administered and used, collect ongoing evidence about pass/selection rates for test takers from the different linguistic groups.

6.7. Conduct quality control studies separately for each stage of the assessment process to serve as a basis for ensuring score comparability.

6.7.1. If the test is to be administered to all the linguistic populations in the same language (e.g., the reference language), conduct studies to help ensure score comparability between the populations.

Guideline 7: Test Preparation

7.1. To aid familiarize all linguistic groups with the test items and format, the test user should provide test takers from all linguistic and cultural groups with approved practice, sample, or test preparation materials consistent with the recommended practice for the test. All materials (including test instructions) should be adapted in ways that do not change the assessed construct. In addition, descriptions of materials that are reflective of other acceptable accommodations should be provided, as relevant.

7.2. The test publisher should provide a complete description of the test, its characteristics, and its purpose. This is particularly important in the assessment of test takers from different linguistic or cultural groups because it provides an opportunity to minimize potential differences arising out of differences in the understanding of item types or item formats that may be unfamiliar to some test-taker populations.

7.2.1. The test preparation materials should contain information about the specifics of the test. This includes timing information, the numbers of sections in the test, samples of each of the various item types that will appear in the test across each of the content areas, and how and when scores will be provided. This will enable all populations to become familiar with the content and format of the test, including instructions for each item type, test mode, test length, timing, and scoring (including information about penalty scoring, if applicable).

7.2.1.1. If possible, prepare a list of test practice strategies for test takers who, because of their linguistic background, may be unfamiliar with preparing for certain kinds of tests. These may include suggesting setting up study groups (comprised of test takers from the same, from other, or from mixed linguistic groups) and the use of ancillary study resources.

7.3. If the test is to be computer-delivered, indicate whether it will be adaptive or linear (fixed), and explain what that means in practice. For adaptive tests, explain that the difficulty of the test is based on how well the test taker performs on earlier test items and that the level of difficulty of subsequent items may increase quickly to try to match the test taker's proficiency level.

7.4. Where possible, provide the correct answers and rationales for the correct answers for each item in the sample test to provide test and item familiarization to test takers from all linguistic groups.

7.5. Explicitly clarify the differences between test preparation approaches (such as coaching and instruction) and identify the one deemed acceptable for the test.

Glossary of Terms and Definitions

Accommodations – Adaptations that are made to the design of an assessment or its administration that do not alter the measurement of the underlying construct or the interpretations of the scores on that assessment.

Adaptations (Adapted Tests) – Any changes that are made to the design of an assessment in terms of content, format, or test administration to increase access to the material on the assessment for cultural or linguistic groups that may differ from the mainstream population. There are implications for modifying a test or its administration, which will have implications on the score interpretation(s). These implications should be considered jointly by the test developer and the test user(s).

Adaptive test – An adaptive test, normally administered by computer, is one in which an algorithm determines whether test items that are less difficult, more difficult, or are of the same difficulty are administered to a test taker, based on that test taker's performance on previous items. The aim of these tests is to provide better measurement of each test taker's ability.

Analytic (trait) scoring – A method of essay scoring in which each critical dimension (such as grammar, the quality of the argument, writing style, etc.) is judged and scored separately, and the resulting trait scores are combined for an overall score. This is in contrast to holistic scoring.

Benchmarks – Also known as **anchor responses**. Examples include preselected essays used as examples of the different score points on a rubric, and used to train and calibrate raters.

Central tendency – When raters avoid the extreme categories of a rating scale and prefer categories near the scale's midpoint.

Coaching – Short-term approaches to test preparation, such as easy test-taking strategies or quick fixes to help boost scores.

Comparability – In assessment, score comparability indicates the extent to which similar inferences can be made based on scores across similar assessments. Interpretations of scores on adapted assessments for linguistically/culturally diverse populations should carry the same meaning as the original assessments from which they were derived because both assessments should be testing the same construct(s).

Consistency – Degree to which critical features of tests (e.g., items, raters, testing time) are comparable across test conditions.

Construct – The knowledge, skill, ability, or attribute that a test aims to measure. Constructs are not directly observable but are latent.

Criterion-referenced tests – A test whose standardized scores are based on a predetermined set of criteria (or standards) of performance and not against other test takers' performances.

Cut scores – A preselected point (or points) on the scale of an assessment used to distinguish between classes of test takers. A single cut score may be used to classify a test taker as

possessing certain characteristics based on their performance on a test, for example, demonstrating minimal competence in knowledge or ability of a particular construct. Multiple cut scores may be used on an assessment to classify test takers into different predetermined categories based on a group of standards.

Dialect – A form of a language that is used in a specific region or by a particular social group.

Differential item functioning (DIF) – When groups of examinees of *equal ability* select the correct item option at different rates when compared to the reference test-taker group. These differences may be due to construct-relevant or construct-irrelevant sources. Two methods for identifying uniform DIF are commonly used in large-scale operational assessments: (1) **Mantel-Haenszel** –in large focal and reference groups or (2) **Standardization** – when one or both groups are small. For more information, see Dorans and Holland (1992) or Osterlind and Everson (2009). To investigate both uniform and non-uniform DIF within the Confirmatory Factor Analysis (CFA) framework, the **Likelihood Ratio test** or a **Wald Test** for differences of discrimination and difficulty parameters are preferred.

Discrepant scores – Depending on the score scale range, when the scores that are assigned by two raters differ by a particular number of points (e.g., more than one point).

Fairness – The notion of fairness in assessment is the idea that the inferences that are drawn from performances on an assessment are the same, regardless of a test taker's background or membership to a particular population. With regard to cultural and linguistic diversity, this implies that if an individual is administered an adapted assessment, that individual's performance would validly demonstrate their knowledge or ability of the targeted construct.

Generalizability theory (G-theory) – allows for the recognition of multiple sources of measurement error, estimating the magnitude of each source separately, and providing evidence for minimizing the measurement error of the assessment.

Halo effects – Type of bias in which a rater provides similar ratings on conceptually distinct criteria or performance aspects (e.g., a rater's general impression of a test taker or test taker's response similarly affects each criterion on the scoring rubric).

High-stakes (vs. low-stakes) assessments – **High-stakes assessments** have implications for important decisions about the test taker in terms of admissions decisions, ranking, scholarship, or diagnosis. **Low-stakes assessments** are meant to assess an individual at a point in time to determine future decisions by a teacher, professor, or psychological practitioner (e.g., learner feedback, course assignment).

Holistic scoring – When essays are awarded a *single*, overall score by a rater that reflects all aspects of an essay, according to a description of that score point in the assigned rubric.

Instruction – long-term test preparation approach that intends to improve knowledge and skills.

Instrument bias – When population performance differences between the original test and adapted test forms are revealed that are due to construct-irrelevant sources. This may be due to differential familiarity with stimulus material, differential familiarity with response procedures,

differential response styles (such as cultural response sets, e.g., a differential preference for self-disclosure), or differential social desirability.

Inter-rater reliability – Refers to the consistency of scores given by different raters, or agreement of scores between two or more raters.

Leniency (severity) – The overall tendency of a rater to score essays either too stringently (strictly, severely) or too relaxed (leniently).

Linear test – In a linear test, all items of the test are delivered to all test takers in the same order, regardless of a test taker's performance on previous test items.

L1 – Refers to the home/first language(s) of a test taker.

L2 – Refers to the second language or foreign language of a test taker.

Many-facet Rasch measurement (MFRM) – An approach that allows studying rater effects such as severity/leniency, halo, and central tendency, as well as examining the utility of the rating scales and the presence of differential rater functioning, such as when some raters are biased against groups of test takers, in particular against cultural or linguistic groups.

Modular test design – A test that is comprised of multiple sections, some or all of which can be interchanged with other equivalent test sections.

Multistage Adaptive Testing – Similar to adaptive testing, but instead an algorithm determines which *groups* of test items are administered to a test taker, based on that test taker's performance on previous items that were less difficult, more difficult, or of the same difficulty.

Norm-referenced test (normative score) – A test whose standardized score reports how well test takers perform against the performance results of a statistically selected group of test takers from a normally distributed group.

Objectivity – Objectivity is the inherent quality of an object in itself, unrelated to any speculative approach; this attribute acts in favor of equity and fairness of testing. Among its properties are: (a) absence of bias in interpretation and decision-making, (b) focus on impartiality for test raters free from individual assumptions and values, (c) distinction between two contrasting or even conflicting ideas or theories based on the exact definition of the object (Gaukroger, 2012). Objectivity is an attribute not limited to the use of multiple-choice questions or other closed form items.

Official language – The language or one of the languages that is approved by the government of a country, for use in legal and official (government) documents, is taught in schools, and is used in the legal system.

Polytomous items – Test items that have more than two response or score categories.

Rangefinders – Prior to a scoring session, essays are selected as benchmarks (see definition above) and then are used for the training and calibration of the individuals who will be assigning scores to each essay from the pool of essays.

Rater bias – When the scores that a rater gives to test takers' responses change because of some aspect of the assessment situation that is not relevant to the measured construct; for example, when raters assign a more severe rating to members of only some groups.

Rater effect – When a rater(s) systematically assigns scores to essays that may not necessarily be an objective reflection of the rubric. The types of rater biases are central tendency, halo, leniency, and severity effects (all defined in this glossary).

Raters – Individuals who score essays and other assessments. They also are referred to as judges, readers, markers, graders, scorers, or checkers.

Rating scale – The entire range of possible scores for an assigned item, such as an essay, sometimes accompanied by performance descriptors of traits or behaviors assessed at some or all of the score points.

Raw scores – The total number of items that were answered correctly on an assessment or the sum of item scores when polytomous item responses are used. Variations on the way the raw test score is derived from item scores are possible (for example, weighted sum can be used); however, raw scores are not further adjusted or transformed in any way.

Read-aheads (Seeded responses) – Used as a quality check to ensure rater calibration. This is accomplished by having raters score essays that were previously scored by the scoring leader to ensure that their scores have not drifted.

Read-behinds – Second scores on essays given by expert staff based on a random sample of responses already scored by their scoring team members.

Readers – See raters.

Reliability – Fundamentally, refers to the measurement precision of the test. Different methods for estimating test reliability exist (e.g., inter-rater reliability, test-retest reliability, parallel forms reliability). *Inter-rater reliability* indicates the consistency (agreement) of scores between different raters for test takers' answers on constructed-response items. *Test-retest reliability* indicates the consistency of scores for a test given on multiple occasions. *Parallel forms reliability* indicates the consistency of scores between different test forms of the same assessment.

Rubrics – A guide that contains criteria for the scoring of test takers' essays or other constructed responses.

Scores – Another word for ratings, grades, or marks.

Sociocultural – Used to describe the combination of social and cultural factors.

Sociolinguistics – The study of how the use of language is impacted by a wide range of social situations, including differences between groups based on regions, social classes, gender, and occupations.

Source and targeted languages and cultures – A source language is the original language of a test. A targeted language is the language to which the test is to be translated.

Speededness – A characteristic of an assessment when the test taker's score not only depends on the correctness of item responses but also the rate at which those items are completed. In terms of fairness, a speeded test may disadvantage L2 test takers, particularly when language proficiency is not the targeted construct of the test.

Standardized scale scores – A transformation of a test taker's raw score to a standardized range of scores (scale). This allows for a meaningful comparison of all test takers to a normative population, and between different test forms for the same assessment.

Subscores (subscale scores) – A score for a particular construct that is part of the overall composite test score.

Task-specific scoring – Refers to scoring the performance on a task with respect to a separate feature or set of features particularly relevant to that task. This can be realized by a holistic scale (primary trait scoring) or by a task-specific set of analytic scales (multiple-trait scoring).

Test administration manual – Contains policies and procedures for test administrators, proctors, and other people who will administer a given test. The manual should provide instructions to ensure testing conditions are standardized in terms of the handling of test materials, timing of the test, the test environment, proctoring, administration of accommodations, and procedures regarding disruptions or suspected cheating.

Test proctor – The person responsible for administering an assessment to test takers and ensuring that all test administration procedures are properly followed. Test proctors are also responsible for monitoring the test takers to answer any procedural questions they may have. They also ensure that test takers do their own work and do not copy answers from other test takers.

Test takers – Another word for examinees, candidates, or respondents.

Test user – The user of a test are the professionals from an organization that select an instrument to measure specific attributes for a specific purpose. They are responsible for properly interpreting meaning from test takers' scores and making inferences about the evidence that those scores represent. They are also responsible for ensuring that an assessment is properly administered under standardized conditions.

Translation – The process of converting an assessment from one language into another language (or languages) such that the same construct is measured, the difficulty of each translated item is the same as the original, and the scores on both tests are comparable, that is, the same interpretations can be made about the scores on both tests.

Validity – Validity of an assessment refers to the degree to which theory and empirical evidence support the intended meaning and use of assessment outcomes.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington D.C.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and Standardization. *ETS Research Report Series*, *1992*(1), i-40.

Eckes, T. (2015). Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments (2nd ed.). Frankfurt am Main, Germany: Peter Lang.

Educational Testing Service (2015). *ETS Guidelines for Fair Tests and Communications*. Retrieved from: https://www.ets.org/s/about/pdf/ets_guidelines_for_fair_tests_and_communications.pdf

Educational Testing Service (2009). *ETS International Principles for Fairness Review of Assessments - A Manual for Developing Locally Appropriate Fairness Review Guidelines in Various Countries*. Retrieved from: https://www.ets.org/s/about/pdf/fairness_review_international.pdf

Educational Testing Service (2009). *Guidelines for the Assessment of English Language Learners*. Retrieved from: http://www.ets.org/s/about/pdf/ell_guidelines.pdf

Elosua, P. (2016). Minority language revitalization and educational assessment: Do language-related factors impact performance? *Journal of Sociolinguistics*. 20(2), 212-228.

Engelhard, G., & Wind, S. A. (2018). Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments. New York, NY: Routledge.

Gaukroger, S. (2012). *Objectivity. A very short introduction*. Oxford: Oxford University Press.

Haugen, E. (1966). Dialect, Language, Nation. *American Anthropologist*, 68(4), 922–935. Retrieved from http://www.jstor.org/stable/670407

International Test Commission (2017). *Guidelines for Translating and Adapting Tests, 2nd edition*. Retrieved from: https://www.intestcom.org

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). Assessing performance: Designing, scoring, and validating performance tasks. New York, NY: Guilford.

Luykz, A., Lee, O., Mahotiere, M., Lester, B., Hart, J., & Deaktor, R. (2007). Cultural and home language influences in children's responses to science assessments. *Teachers College Record*, 109(4), 897-926.

Oakland, T. (2016). Testing and assessment of immigrants and second-language learners. In: Leong, F. et al. (Eds.). *The ITC International Handbook of Testing and Assessment*. Oxford University Press.

Oliveri, M. E., Lawless, R. R., & Mislevy, R. J. (2018*). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments*. (Manuscript in press).

Oliveri, M.E., Lawless, R., & Young, J. (2015). *A validity framework for the use and development of exported assessments*. Princeton: ETS Office of Professional Standards Series. Retrieved from: https://www.ets.org/s/about/pdf/exported_assessments.pdf

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Los Angeles, CA: Sage.

Sireci, S. G., Harter, J., Yang, Y., & Bhola, D. (2003). Evaluating the equivalence of an employee attitude survey across languages, cultures, and administration formats. *International Journal of Testing*, 3, 129-150.

Survey Research Center. (2016). *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved June 21, 2018 from http://www.ccsg.isr.umich.edu/.