CAL CENTER FOR APPLIED LINGUISTICS

CAL Commentary
July 2018

Ensuring Fairness in Language Proficiency Assessments: Q&A

Caitlin Gdowski & Keira Ballantyne

Language proficiency assessments may be used for multiple purposes, some of which have high stakes outcomes for the examinees or for society (Bachman & Palmer, 1996). Outcomes that impact individual examinees might include professional opportunities, immigration statuses, and eligibility for educational services or opportunities. Because these outcomes are so significant for the individuals involved, it is critical that the test be a trusted source of information about a person's language proficiency. Test items which inadvertently favor or disfavor individuals may introduce bias into the testing process.

Therefore, it is critical that all examinees who take language proficiency assessments have the same opportunities to demonstrate their language proficiency skills. Language proficiency assessments are trustworthy when they are valid and reliable tools.

This short Q&A aims to provide insight into how test development and psychometric researchers at CAL identify biased items and prevent sensitive topics from appearing on CAL's language proficiency assessments. These considerations help ensure that CAL's assessments are fair, valid, and reliable.

Q: What is bias?

A: Bias in an assessment results in systematically influenced scores based on test taker characteristics that are not relevant to the ability being measured (McNamara & Roever, 2006). A biased item on a language proficiency test is one on which examinees with similar language ability perform differently for reasons unrelated to their language proficiency. For example, an examinee who has never skied should not be asked to explain the steps for learning how to ski. We want to be sure that test items are fair and accessible to every student regardless of their background. Items on language proficiency assessments should measure language proficiency skills, instead of a student's knowledge of a content area or an experience (Bachman & Palmer, 1996). All items on CAL's language proficiency assessments are evaluated by trained reviewers who examine test item text and images for potential bias. Multiple aspects of examinees' characteristics should be considered when

> *"Bias is the presence of some characteristic of an item that results in the differential performance of two individuals of the same ability but from different ethnic, sex, cultural, or religious groups." (Hambleton & Rodgers, 1995, p.1)*

developing and reviewing test content. Some considerations include: gender, age, religion, socio-economic status, home language, race, ethnicity, culture, family type, age-of-arrival to the U.S., geographic region within the U.S., and disabilities.

Q: What are sensitive topics?

A: Sensitive topics are those that may elicit a negative emotional response from test takers and negatively impact their ability to accurately demonstrate their language proficiency ability (McNamara & Roever, 2006; Zieky, 2006). For example, personal experiences as well as ethnic, cultural, or religious beliefs or practices may influence students' reactions to a topic. We want to ensure that sensitive topics do not appear on CAL's language proficiency assessments because an upsetting topic may prevent examinees from accurately demonstrating their language proficiency due to a negative or distracting response or association to a test item. For this reason, CAL collaborates with individuals familiar with the test population to identify sensitive topics to avoid for a given assessment. Test-specific lists of sensitive topics guide the development and review of assessment materials, such as prompts, passages, items, and graphics.

> *"Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups…" (AERA, APA, & NCME, 1999, p.82)*

Q: Which activities in the item development process help ensure fairness of assessments?

A: Language testing specialists are trained to develop test items that avoid biased or sensitive topics, and external reviewers are also empaneled during item development in order to ensure that CAL's language proficiency assessments are fair. Two main activities that help ensure fairness are Topic Generation and Bias and Sensitivity Review.

Topic Generation

- Focus groups are conducted with experienced educators who are familiar with the assessment's target population to discuss and generate item topic ideas.
- Item quality is evaluated considering
    - connection to relevant standards,
    - robustness,
    - age of test takers,
    - accessibility, and
    - sensitivity.

- Overall, topics and graphics should reflect
  - diverse cultures, ethnic and socio-economic groups, regions, and individuals with varying physical abilities;
  - balanced gender roles; and
  - positive situations, language, and images.

Bias and Sensitivity Review

- All new items undergo an external bias and sensitivity review, also called a fairness review.
- The goals of the bias and sensitivity review are to
  - review newly-created language proficiency test items for potential bias and sensitivity issues, and
  - identify and discuss issues and possible solutions.
- In this process, reviewers
  - use their experiences as educators or test development professionals to identify any potential bias or sensitivity issues,
  - provide their unique perspectives that represent their students and contexts,
  - contribute to creating a positive group dynamic where potentially-problematic items are discussed and solutions are proposed, and
  - provide useful feedback to improve the test content.

Q: How does quantitative analysis of assessment items contribute to fairness in language proficiency tests?

A: Differential Item Functioning (DIF) analysis is a quantitative method of examining items for potential bias.

Differential Item Functioning Analysis

- DIF analysis is a statistical procedure which compares the performance of matched examinees on a given test item by subgroup variable, such as gender or ethnicity. In DIF analysis, first all examinees are assigned into distinct ability bands, by selecting groups of examinees who have equal estimates of overall ability. The performance of these groups of matched examinees is examined item-by-item to detect any items for which there appears to be systematic differences in item difficulty by subgroup, such as gender or ethnicity.
- Any item which meets the industry standard criteria for a DIF level of concern (C-level DIF) is examined by a

Differential Item Functioning (DIF) analysis provides quantitative evidence of fairness.

Qualitative and quantitative evidence of the steps that ensure a bias-free assessment are made available in technical reports of assessments.

content expert panel (Zwick, 2014). Panels are balanced by gender and by language background, incorporating individuals with diverse linguistic and cultural backgrounds. The panel recommends either that the item should not be used due, to concerns over bias, or finds that the item is appropriate for operational testing.

- If the panel recommends the item is not appropriate for operational use, possible response actions include
  - removing the item from the pool of items available for new form selection,
  - removing the item from operational scoring, or
  - modifying the item to remove bias concerns and re-field testing the item.

References

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

Bachman, L.F. & Palmer, A.S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Hambleton, R.K. & Rodgers, H.J. (1995). Item bias review. *Practical Assessment, Research & Evaluation*, *4*(6), 1-3.

McNamara, T.F. & Roever, C. (2006). *Language Learning Monograph Series. Language Testing: The Social Dimension.* Malden, MA: Blackwell Publishing.

Zieky, M. (2006). Fairness Review in Assessment. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of Test Development* (pp.359-376). Mahwah, NJ: Lawrence Erlbaum Associates.

Zwick, R. (2014). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series, 2012*(1), i-30.

About the Authors

Keira Ballantyne manages the Psychometrics and Quantitative Research team at CAL. Dr. Ballantyne holds a PhD in linguistics (University of Hawai'i at Mānoa, 2005) and has over ten years of experience in research, policy, and practice with culturally and linguistically diverse (CLD) K-12 student populations. Prior to her work at CAL, Dr. Ballantyne was a Research Scientist at the Center for Equity and Excellence in Education, and prior to that, Director of the National Clearinghouse for English Language Acquisition, both at The George Washington University. Dr. Ballantyne has produced research and practitioner-focused publications spanning broad topics in the education of culturally and linguistically diverse learners, including on early childhood education for dual language learners, on teacher preparation, on high school graduation and success for CLD learners, on CLD learners with learning disabilities, and on formative assessment of literacy. She has also served as a teacher educator and an ESL instructor for adult immigrant students in the United States.

Caitlin Gdowski is a Language Testing Specialist on the Test Development Team at CAL. She primarily develops content and works with external consultants to refine content for ACCESS for ELLs. She contributes to other assessment programs at CAL by developing new content, reviewing pre-operational forms, and rating test-taker responses. Prior to working at CAL, Ms. Gdowski was a member of the Assessment Division at the University of Michigan English Language Institute (now Michigan Language Assessment). Ms. Gdowski taught EFL in Kyrgyzstan as a Peace Corps Volunteer. She holds a Master's degree in English Language Teaching, specializing in Testing and Assessment, from the University of Warwick as well as a B.A. in Linguistics and a B.A. in Brain, Behavior, and Cognitive Sciences from the University of Michigan.

About CAL

The Center for Applied Linguistics (CAL) is a non-profit organization founded in 1959. Headquartered in Washington DC, CAL has earned an international reputation for its contributions to the fields of bilingual and dual language education, English as a second language, world languages education, language policy, assessment, immigrant and refugee integration, literacy, dialect studies, and the education of linguistically and culturally diverse adults and children. The mission of the Center for Applied Linguistics (CAL) is to promote language learning and cultural understanding by serving as a trusted resource for research, services, and policy analysis. Through its work, CAL seeks solutions to issues involving language and culture as they relate to access and equity in education and society around the globe.

CAL
CENTER FOR APPLIED LINGUISTICS
www.cal.org

Promoting Access, Equity, and Mutual Understanding
Among Linguistically and Culturally Diverse People Around the World