Avoiding bias in language test development

Benjamin Kremmel

Language Testing Research Group Innsbruck
University of Innsbruck

ALTE Ljubljana, November 2019
Overview

• What is bias?
  • Examples

• Some sources of bias

• How can we avoid bias
What do we mean when we say a test is biased?
Bias...

• is “systematic error that disadvantages the test performance of one group” (Shephard, Camilli, & Averil, 1981)

• the internal properties of a test disadvantage or negatively affect particular subgroups of test takers
Examples (Roever, 2007)

• Imp 12: Mike is trying to find an apartment in New York City. He just looked at a place and is telling his friend Jane about it.
  • Jane: “Is the rent high?”
  • Mike: “Is the Pope Catholic?”

• What does Mike probably mean?
  • He doesn't want to talk about the rent.
  • The rent is high.
  • The apartment is owned by the church.
  • The rent isn't very high.
• „You rent a house through an agency. The heating system has stopped working. You phoned the agency a week ago, but it has still not been mended. Write a letter to the agency. Explain the sitation and tell them what you want them to do about it.“
Examples (TOEFL listening prep Kit, from Djiwandono, 2006)

(man) I'm taking up a collection for the jazz band. Would you like to give?
(woman) Just a minute while I get my wallet
(narrator) What will the woman probably do next?

29. a. put some money in her wallet
    b. buy a band-concert ticket
    c. make a donation
    d. lend the man some money
Examples  (TOEFL listening prep Kit, from Djiwandono, 2006)

(man) Can you go over my notes with me? I'll never understand all these chemistry experiments.
(woman) You know, review sessions are being held every night this week. They are supposed to be good.
(narrator) What does the woman imply the man should do?

16.  a. make a copy of his notes for her  
b. ask his professor for help  
c. attend the review sessions  
d. go to the chemistry lab this evening
Bias...

• is “systematic error that disadvantages the test performance of one group” (Shephard, Camilli, & Averil, 1981)

• the internal properties of a test disadvantage or negatively affect particular subgroups of test takers

• E.g. “gender, instructional experience, background knowledge associated with particular disciplinary areas, first language background, and ethnicity of particular test-taker groups” (Elder, 2012)
Bias...

- threatens the validity of interpretations
- jeopardizes test fairness
Bias and construct-irrelevance

Construct under-representation

CONSTRUCT

VALIDITY

construct-irrelevant variance

TEST

Watts (n.d.)
Bias and construct-irrelevance

Watts (n.d.)
Fairness (Kunnan, 2008)

- Validity
- Absence of bias
- Access
- Administration
- Social consequences

“Test developers should strive to make tests that are as fair as possible for test takers of different races, gender, ethnic backgrounds, or different handicapping conditions.”
Caveat

• Differences in group performance in themselves do not necessarily indicate the presence of bias, since differences may reflect genuine differences between the groups on the ability in question (Bachman, 1990, p. 270)

• Bias occurs when these differences are not logically related to the ability in question!
What are potential sources of bias?
Range of forms (Bachman, 1990)

- Misinterpretation of test scores
- Sexist or racist content
- Unequal prediction of criterion performance
- Unfair content with respect to the experience of test takers
- Inappropriate selection procedures
- Inadequate criterion measures
- Threatening atmosphere
- Conditions of testing
1. *Content or language variety*: This type of bias refers to content or language or dialect that is offensive or biased to test takers from different backgrounds. Examples include content or language stereotypes of group members and overt or implied slurs or insults (based on gender, race and ethnicity, religion, age, native language, national origin, and sexual orientation) or choice of dialect or variety that is biased to test takers.

2. *Group performance*: This type of bias refers to difference in performances and resulting outcomes by test takers from different group memberships. Group differences could occur among salient groups (e.g., gender, race and ethnicity, religion, age, native language, national origin, and sexual orientation) on test tasks and subtests.

3. *Standard setting*: This type of bias refers to standard setting in terms of the criterion measure and selection decisions and how these decisions affect different test taking groups.
Sources of bias (I) (Reynolds & Suzuki, 2012)

• *Inappropriate content.* Tests are geared to majority experiences and values or are scored arbitrarily according to majority values. Correct responses or solution methods depend on material that is unfamiliar to minority individuals.

• *Inappropriate standardization samples.* Minorities' representation in norming samples is proportionate but insufficient to allow them any influence over test development.

• *Examiners' and language bias.* White examiners who speak standard English intimidate minority examinees and communicate inaccurately with them, spuriously lowering their test scores.
Sources of bias (II) (Reynolds & Suzuki, 2012)

- **Inequitable social consequences.** Ethnic minority individuals, already disadvantaged because of stereotyping and past discrimination, are denied employment or relegated to dead-end educational tracks. Labeling effects are another example of invalidity of this type.

- **Measurement of different constructs.** Tests largely based on majority culture are measuring different characteristics altogether for members of minority groups, rendering them invalid for these groups.

- **Differential predictive validity.** Standardized tests accurately predict many outcomes for majority group members, but they do not predict any relevant behavior for their minority counterparts. In addition, the criteria that tests are designed to predict, such as achievement in White, middle-class schools, may themselves be biased against minority examinees.

- **Qualitatively distinct aptitude and personality.** This position seems to suggest that minority and majority ethnic groups possess characteristics of different types, so that test development must begin with different definitions for majority and minority groups.
Rater bias

Language testing research has identified biased rating patterns on the part of

• novice raters vs experts (e.g., Weigle, 2002)
• language vs discipline specialists (e.g., Brown, 1995)
• native vs non-native speakers (e.g., Kim, 2009)
Rater bias

• NS more severe than NNS (Hill, 1997)
• NNS more severe than NS (Fayer & Krasinski, 1987)
• Raters more severe with participants from same L1 (Harding & Griffiths, 2016)
• Familiarity with speaker’s L1 influences pronunciation assessment (Carey et al., 2011)
• “L2 familiarity” → more lenient rating (Winke et al., 2012)
Further sources of bias

- Halo- and other rating effects
- Test delivery / lack of accommodations
- Representativeness among teachers/testers/course book authors/item writers (e.g. gender)
- …
How Can Bias in Assessment be Avoided?

Group Task ([https://www.k-state.edu/ksde/alp/activities/Activity4-2.pdf](https://www.k-state.edu/ksde/alp/activities/Activity4-2.pdf)):

Choose 1-2 assessments that are administered to students in your classroom/school/district/region. Individually, consider whether or not bias is present in each assessment by addressing each of the items in the table on the following page. Be certain to highlight those items that yield a “No” answer.
1. Was bias present in any of the assessment items you reviewed?
2. What needs to be done to correct the items?
3. How can you avoid test bias in future assessments?
4. What could be the harm to your students if you don’t address the issue of avoiding test bias?
5. Compare the results of your reviews with other group members.
6. Why is doing this type of bias review more effective if a group of teachers works together to review a test – rather than one individual?

<table>
<thead>
<tr>
<th>Bias Review:</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are there any test items that:</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Contain language that is not commonly used or has different connotations in different parts of the state or country, or in different cultural or gender groups?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portray anyone in a stereotypical manner?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contain any demeaning or offensive materials?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have any religious references?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have references that mean different things to different cultures?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assume that all students come from the same socioeconomic or family background?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contain information or ideas that are unique to the culture of one group AND this information or idea is not part of the content standards?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measure membership in a group more than measure a content objective?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Put up barriers preventing any group of students from demonstrating their knowledge and abilities?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portray a group unfavorably or in a stereotypical manner?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contain language or symbolism that can be interpreted in an offensive or emotionally charged way to a person or group?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What would a list of recommendations to avoid test bias look like for your assessment context?
Reduce bias during test design by ...

(1) using heterogeneous sets of test writers and editors;
(2) taboo topic lists and awareness-raising in item writer and rater training
(3) avoiding any items that assume general knowledge not part of the relevant program;
(4) having test materials (incl. pictures!) reviewed by members of minority groups;
(5) examining item data from tryouts or prior administration separately by group (e.g. DIF analysis)

(adapted from Green, 1979)
Preempting offensive content (Elder, 2012)

• Scrutiny of test content during item development and bias review committees

• Fairness guidelines for item developers (e.g. ETS, 2009) to avoid unfair content or images (e.g. (gender) stereotyping, inappropriate labeling of ethnic groups, reference to distressing incidents or controversial topics, use of graphs that may be unduly difficult for candidates with limited numeracy, or specialized words that might disadvantage those without relevant technical knowledge
Reduce bias during test administration by ...

(1) using examiners familiar to the examinees;
(2) making testing situations similar to the learning situation;
(3) providing repeated practice tests with feedback;
(4) keeping examiners (ethnically) heterogeneous

(adapted from Green, 1979)
Reduce bias during test **scoring** by ...

1. using only objectively scorable measures;
2. training personnel to make legitimate generalizations from test scores;
3. specifying the intended use of scores.

(adapted from Green, 1979)
Fulcher’s recommendations (2010)

• Bias/Sensitivity review

• Check that items do not contain references or materials likely to lead to bias against a certain subgroup.

• Cultural sensitivity (representative review panel)

• Identify DSIs (designated subgroups of interest) and for each DSI ask if any member is likely to suffer because the content is beyond their educational or cultural experience, whether it is inflammatory, offensive or portrays some DSIs stereotypically.
Fairness review guide (ETS, 2009)

• Measure the important aspects of the relevant content
• Avoid irrelevant cognitive barriers to the performance of test takers
  • Language, topics, translation, contexts, religion
• Avoid irrelevant emotional barriers to the performance of test takers
  • Advocacy, sensitive topics, stereotypes, appropriate terminology, representation of diversity
• Avoid irrelevant physical barriers to the performance of test takers
  • E.g. accommodations

→ Develop guidelines, establish procedures, train users, monitor and revise guidelines, conduct validation research
• What characteristics define the groups that should receive special attention in the development of your guidelines? For example, in some countries the type of school a test taker attended could be a relevant factor.

• What level of vocabulary and syntax is acceptable for the tests you are developing? How would you describe “accessible language” for item writers to use? What aspects of language should item writers avoid unless language is the intended focus of measurement?

• What aspects of specialized knowledge that are not the point of measurement are likely to be irrelevant cognitive barriers in your country?

• Which topics would be of concern in translated tests used in your country? What additional topics would be of concern?
Fairness review guide (ETS, 2009)

• What contexts are likely to be appropriate for the tests you are developing? Are there contexts that should be avoided?

• How should religion be treated in tests in your country? Is there some knowledge about religion that all test takers are assumed to have, or should religion be avoided unless it is the focus of measurement?

• What topics are so divisive in your country that advocacy of one side or the other should be avoided in tests unless required for valid measurement?

• What topics are so sensitive in your country that it is best to avoid them in tests unless they are required for valid measurement? For example, in some countries criticism of the royal family must be avoided.
• In your country, what topics must be handled with care because they are likely to present emotional barriers to the performance of test takers?

• What stereotypes should be avoided in tests in your country?

• Which groups may be of concern regarding appropriate terminology in your country? For each group, describe the terminology that is appropriate to identify the group in your country.

• Which groups should be represented in the tests in your country? Approximately what proportion of items that mention people should be allocated to representing diverse groups?
Fairness review guide (ETS, 2009)

• Are special guidelines needed for K-12 tests in your country? If so, which topics should be avoided unless they are required for valid measurement?

• Physical barriers are likely to be very similar across countries because they are caused by sensory and motor problems that can affect any human being rather than by cultural, linguistic, or other issues that vary across countries. What physical barriers are of concern in your country?

• For use in your country, which procedures should be adopted? Which modified? Which rejected? Are additional procedures required?

• Which of these factors should be included in the training of test developers in your country? Should any factors be added?
ALTE’s recommendations (2011)

• Cultural bias (background, age)?
• Do not choose texts that may be biased (culture, gender, age, etc.)
• Topic list (e.g. local customs)
Detecting item bias

Item bias occurs when items unfairly favour or disfavour certain groups of test takers of the same ability. For example, an item may be easier for female test takers than male test takers, even though they are of equal ability. This is unfair because the aim of the test is to measure differences in language ability and not in gender (see Section 1.4).

Care should be taken when diagnosing bias, however, as not all differences between groups are unfair. Learners with a particular L1 may find an item more difficult than learners of the same ability in another group due to differences between the mother tongue and the target language. In the context of measuring language proficiency, this must be accepted as part of the nature of proficiency in the target language, not a problem in measuring it.

One approach to minimising bias is to use a Differential Item Functioning (DIF) methodology to detect possible bias so that it may be investigated further. This involves comparing the responses of groups of test takers who are equally able. For example, if the test is intended to be for adults of all ages, the performance of younger and older adults with approximately the same ability (according to the test) can be compared. Analysis based on IRT is well suited to do this.
Final thoughts

Can/should bias be eliminated completely?

• E.g. Jensen (1980):
  • (a) the egalitarian fallacy, that all groups were equal in the characteristics measured by a test, so that any score difference must result from bias;
  • (b) the culture-bound fallacy, that reviewers can assess the culture loadings of items through casual inspection or armchair judgment;
  • (c) the standardization fallacy, that a test is necessarily biased when used with any group not included in large numbers in the norming sample
Thank you!

Benjamin.kremmel@uibk.ac.at
@benjaminkremmel
References


References (cont’d)