

"That's not fair!"

Fairness, bias, and differential item functioning in language testing

Carsten Roever

The University of Melbourne

carsten@unimelb.edu.au

- Aspects of fairness in testing (based on Kunnan, 2000; Shohamy, 2000)
 - Construct – adequate measurement of the attribute under investigation
 - Content – domain content adequately covered
 - Standardization – same conditions, materials, scoring for all test takers
 - Unbiased Items – items do not advantage or disadvantage sub-groups
 - Test Consequences / Score Use – scores are used to make appropriate decisions
- Both, Kunnan and Shohamy, also mention societal issues, including equal learning opportunities, equal access to the test, and social impact of testing.
- I will focus on bias at the item level, and not talk about oral proficiency testing or rater issues

History of bias research

- Fairness concerns are often framed in terms of test bias favoring or disadvantaging groups of test takers by gender, native language, ethnicity, or socio-economic status
- Bias received attention as early as 1911, when Binet was forced to revise his earlier contention that his intelligence test measures "mental capacity" after finding that children of high socio-economic status performed far better than children of low socio-economic status (Eels et al., 1951)
- Test scores were obviously influenced by educational background and opportunity to learn, not only by "raw" intellectual capacity

- Eels et al. (1951) summarize the research from the first half of the 20th century and present the optimistic outlook that sources of bias in items can be systematically identified and eliminated
- While there was such early work in intelligence testing, little was done in other areas such as admission testing, licensing exams, never mind language testing

The Golden Rule Settlement

- These issue came into focus through the Golden Rule Settlement in 1984
- In 1976, the Golden Rule Insurance Company sued the Illinois Department of Insurance and ETS for alleged racial bias in the licensing test for insurance agents
- The general passing rate was only 31% and it appeared that the test was "for all practical purposes excluding blacks entirely from the occupation of insurance agent" (Rooney, 1987)
- ETS revised the test, got the lawsuit dismissed, Golden Rule refiled, ETS got the suit dismissed, Golden Rule appealed and got the suit reinstated, with the Circuit Court finding that ETS as the test maker was responsible for the test's effects
- The Circuit Court also found that test makers could be held liable for intentional racial discrimination if they had knowledge that the test was disadvantaging minority test takers but failed to remedy the situation.
- ETS and Golden Rule settled out of court in 1984; ETS agreed to consider any item biased for which correct answer rates for white and African-American test takers differed by more than 15 percent.
- Also, items that fewer than 40% of African-American test takers answered correctly, were flagged as too difficult.
- The test makers agreed that items would be pre-tested, and that future test forms would be constructed to the largest extent possible from unbiased and non-difficult items (subject to content specifications)
- The Golden Rule case increased the urgency for test makers to ensure fairness and equity of all their tests, and for researchers to find ways of detecting bias

Finding bias in tests: Differential item functioning

- Bias is "systematic error that disadvantages the test performance of one group" (Shepard, 1981) or "systematic under- or overestimation of a population parameter by a statistic" (Jensen, 1980)
- The logical first step in detecting bias is to find items where one group performs much better than the other group: such items function differentially for the two groups and this is known as Differential Item Functioning (DIF)
- In all DIF analyses, the test taker population is split into a reference group and a focal group; the focal group is of particular interest, and the suspicion is that they might have been disadvantaged
- DIF is a necessary but not sufficient condition for bias: bias only exists if the difference is illegitimate, i.e., if both groups should be performing equally well on the item
- An item may show DIF but not be biased if the difference is due to actual differences in the groups' ability needed to answer the item, e.g., if one group is high proficiency and the other low proficiency: the low proficiency group would necessarily score much lower
- Only where the difference is caused by construct-irrelevant factors can DIF be viewed as bias
- In such cases, the item measures another construct, in addition to the one it is supposed to measure
- Bias is usually a characteristic of a whole test, whereas DIF is a characteristic of an individual item

Detecting DIF

- The simplest DIF index is based on the differences in correct response rates (p-value) between reference and focal group
- Angoff (1982) proposed the delta plot a.k.a. transformed item difficulty (TID) index that correlated standardized item difficulty estimates for the reference and focal groups; the larger the correlation, the more similar the item difficulties are for the two groups
- However, this method is no longer in use as it confounds item difficulty and discrimination (Angoff, 1993; Camilli & Shepard, 1994): more discriminating items look more difficult than they really are

- Also, the TID method does not match test takers by ability: only where test takers of the same ability level show different likelihood of getting the item right does the item truly function differentially
- An improvement over the TID is the conditional p value (Dorans, 1989; Dorans & Holland, 1993) a.k.a. the standardization procedure, which compares p-values for the reference and focal groups at each score level
- This approach works relatively well even for small sample sizes, particularly when the ability distributions are unequal

TID studies in language testing

- One of the earliest DIF studies in language testing used the delta plot procedure
- Chen & Henning (1985) compared the performance of 77 L1 Chinese and 34 L1 Spanish test takers on the 150 item UCLA ESL placement test
- Their study was replicated (with some alterations) by Kunnan (1990) and Sasaki (1991)
- They computed difficulty indices using the Rasch model for both groups separately and then correlated them; items outside the 95% confidence intervals around the regression line were considered to be showing DIF
- They identified a small number of items, all of which were vocabulary items advantaging the Spanish speakers. One of them (from Sasaki, 1991) was:

This example is *hypothetical*.
 A understand C wrong
 B helpful D imaginary

- Chen & Henning freely admit that their criterion for classifying an item as showing DIF may have been too strict
- Indeed, Sasaki compared the delta plot with Scheuneman's chi square (Scheuneman, 1979), which is also problematic (Angoff, 1993; Baker, 1981), but which found far more DIF than the delta plot

DIF based on Item Response Theory

- IRT techniques are the "gold standard" of DIF detection, as outlined by Thissen, Steinberg, & Wainer (1993)
- The problem with IRT approaches is that they are highly complex mathematically, require specialized computer programs, and need large sample sizes: Camilli and Shepard (1994) warn that sample sizes of less than 500 are likely to yield very unreliable values; a sample size of at least 1000 in the smaller group is probably necessary for stable estimates
- Thissen, Steinberg, and Wainer suggest an approach that first computes item parameters for both groups combined (the "compact" model) and then separately for the reference and focal group, fixing all parameters except for the item under investigation (the "augmented" model)
- Both models are tested for likelihood that they fit the data, and if they fit equally well, no DIF is present because the item behaves the same way for both groups combined as it does for the groups separately
- However, if the augmented model is much more likely to fit than the compact model, there is DIF

IRT DIF in language testing

- Not surprisingly, not many studies of DIF with full-scale IRT have been done in language testing
- Kim (2001) did such a study, using data from Bachman et al.'s Cambridge-TOEFL comparability study (1993)
- Her data set consisted of 1038 test takers who did the SPEAK test, 571 from an Indo-European L1 background (Spanish, French, German), and 467 from an Asian language background (Thai, Japanese, Chinese)
- SPEAK scores are scaled 0-3 and assigned for grammar, pronunciation, and fluency
- This polytomous scoring complicates the use of IRT but allowed Kim to look at differences by score level
- Kim found significant DIF for grammar and pronunciation but not for fluency
- For grammar, the European group was favored at low ability levels but the Asian group was favored at higher ability levels (non-uniform DIF)

- For pronunciation, the European group was consistently favored (uniform DIF) but less so at the highest ability level, i.e., a test taker from the European group needed less ability to get the same pronunciation score than a test taker from the Asian group
- Kim identified the rating rubrics as a possible source of DIF: where the grammar rubric talked about grammatical errors interfering with intelligibility, the European test takers were favored
- Similarly, where the pronunciation rubric talked about "phonemic errors" and "foreign stress and intonation pattern", European test takers were favored

The simpler gold standard: Mantel-Haenszel Odds Ratio

- The Mantel-Haenszel odds ratio is a non-parametric statistic using chi square
- For every score level, it calculates the odds that the reference group will get the item correct, and that the focal group will get it correct
- It then divides reference group odds by focal group odds, showing how much more likely the reference group is to get the item correct
- ETS transforms results from the Mantel-Haenszel procedure into its delta units to classify items as one of three types (Dorans & Holland, 1993):
 - A: negligible DIF, where chi square is not significant or the effect size is less than one delta unit
 - B: intermediate DIF, not in category A or C
 - C: large DIF, where the effect size exceeds 1.5 delta units, and is significantly larger than 1 delta unit
- For test construction, A items are preferred, B items can be used where not enough A items are available and / or due to test specifications, but the use of C items requires an argued case
- Mantel-Haenszel has two major practical advantages: it is conceptually uncomplicated, and it doesn't require highly specialized software; in fact it can be computed with SPSS through Crosstabs with the grouping variable (gender, language) in Rows, the item in Columns, and the matching variable (score) as a Layer
- However, Mantel-Haenszel can be affected by item discrimination and it performs better with large group sizes (Camilli & Shepard, 1994, claim

that it needs similar numbers as IRT procedures but Muniz et al., 2001, show that it functions well with 500 in the smaller group)

- For smaller samples, it may be necessary to collapse score levels but this should be avoided wherever possible

Mantel-Haenszel in DIF research in language testing

- Ryan and Bachman (1992) used data from the TOEFL-FCE comparison study and investigated DIF by gender and L1 (Indo-European vs. Non-Indo-European) with 1426 test takers
- For gender-based DIF, they found no C items, only 4-5% of items were B items, with the rest A items
- For L1-based DIF, the picture was dramatically different: on both tests, 27% of items were C items, with the majority in the vocabulary section
- Another 12% in the FCE and 17% in TOEFL were B items
- However, the number of items advantaging and disadvantaging one group over the other was nearly perfectly balanced in both tests, so it would not actually cause much difference in scores
- Ryan & Bachman point out quite rightly that there is still a possible concern for tests assembled from item banks, because such a test might be biased just by coincidence
- Elder (1996) also used Mantel-Haenszel for her study of L1 background in the Australian Language Certificate

Logistic Regression

- Zumbo (1999) shows how logistic regression can be profitably used for DIF analysis
- Fundamentally, regression shows whether likelihood of a correct response can be predicted from the total score alone (proficiency effect, no DIF), or whether group membership significantly improves the prediction (uniform DIF), or whether the interaction between membership and score level significantly improves the prediction even further (non-uniform DIF): $Y = b_0 + b_1 \text{score} + b_2 \text{group} + b_3 (\text{score} * \text{group})$
- Kim (2001) used logistic regression for her study as well, and while group and the group*score interaction contributed significantly to the regression, their effect sizes were small

- Lee, Breland, and Muraki (2004) used logistic regression for CBT TOEFL writing sections, comparing the performance of 132,941 test takers from IE languages with 121,494 test takers from non-IE languages across 81 different prompts
- Since no test taker took more than one prompt, the score variable was the TOEFL score
- Lee et al found that TOEFL scores almost perfectly predicted essay scores, and that even for the prompts with the largest effect sizes, those effect sizes were still negligible

Some problems with DIF detection

- **Circular reasoning in matching by score level:** since the total score is the matching criterion, no bias will be found if the whole test is biased
- It will just look like the two groups have very unequal ability / proficiency
- **DIF contamination in IRT models:** IRT models calculate item parameters based on all items; if several items are biased, the parameters can be far off
- Scale purification procedures are possible where DIF items are eliminated and item parameters recalculated (Camilli & Shepard, 1994)
- **Definition of Reference and Focal group:** dividing a test taker population by gender, race, or L1 is certainly easy but the value of these binary distinctions is questionable
- And race / gender are mediating variables: it's not race itself that has an effect on scores, but socioeconomic status and educational opportunities that are unequally distributed between races
- It makes a bit more sense with L1 groupings where true similarities between may L1 and L2 help test takers (and false friends can trip them up)
- But there are also cases where amount / quality / focus of L2 education is the actual distinguishing variable
- **No more than two groups are ever compared:** Penfield (2001) shows how multiway comparisons can be done with Mantel-Haenszel

- However, what if an item disadvantages L1 Portuguese speakers in comparison with L1 Japanese speakers but no other significant differences emerge between L1 groups?
- **Unquestioned acceptance of the construct:** what if the construct itself is biased against one of the groups? E.g., what does it mean to have academic English and how much of it is enough?
- This is an issue for validation studies

What to do about DIF?

- DIF items can be altered so that they're less DIF inducing but that presumes that we know what in the item causes the DIF
- Previous studies have shown that men tend to perform better on scientific / practical, sports-related or military content, whereas women perform better on items dealing with human relationships / aesthetics
- Blacks and Latinos perform better than whites on reading passages that deal with minority concerns or contain references to minorities.
- Blacks perform worse than whites on analogy items dealing with science but better on items dealing with human relationships, but this seems to be confounded with whether the item refers to concrete objects (easier for whites) or abstract concepts (easier for blacks).
- Blacks and Latinos perform worse than whites on analogy items that contain homographs.
- Some of these findings may seem almost insultingly stereotypical but they are based on large samples and should only be applied to large samples, not individuals
- Bridgeman & Schmitt (1997) show how item re-writing turned two SAT items classified as C into A items that no longer disadvantaged female test takers:

Vortex:Water::	=>	Whirlpool:Water
(A) volcano:crust		
(B) river:delta		
(C) tornado:air		
(D) geyser:steam		
(E) earthquake:fault		

- Bond (1993) recounts an incident where he and a student devised explanations for DIF in some items only to find that they had identified the wrong items
- When they re-ran their analyses and identified items that actually displayed DIF, their new explanations were the exact opposite of the previous ones

Avoiding DIF (maybe)

- Test makers generally operate a review process to preempt DIF in items
- This process is known as "sensitivity review" or "fairness review"
- It is partially aimed at avoiding construct irrelevant variance, but also at satisfying competing political interests: e.g., ETS avoids mention of evolution in non-biology items but also strives for positive depictions of minorities (see Ravitch, 2003, for discussion of the political influences)
- ETS has published guidelines for fairness review in general (ETS, 2003) and for tests used internationally (ETS, 2004)
- Any item must be reviewed internally by a trained fairness reviewer
- The fairness reviewer can challenge the item and if the developer disagrees with the challenge, a dispute resolution process begins
- Items that are in use for several years should receive a fairness review every 5 years
- For its general fairness review, ETS has six guidelines:
 1. **Treat people with respect:** avoid demeaning language, ethnocentrism, don't degrade or belittle a group, be aware of underlying assumptions (Ramsey, 1993: "The 3,000 Innuvialuit, what the Eskimos of Canada's western Arctic prefer that they themselves be called...")
 2. **Minimize the effect of construct-irrelevant knowledge or skills:** careful with charts / graphs, don't use complex vocabulary where unnecessary, avoid elitism (penthouse, regatta), specialized legal or business terms (junk bonds, subpoena), regionalisms (hoagie, county), specialized sports, tools, transportation terms; general terms are okay
 3. **Avoid controversial, inflammatory, upsetting material:** entirely avoid abortion, genocide, torture, witchcraft; use extreme caution with death, evolution, religion, violence; also be sensitive to cross-cultural issues (depictions of men / women)

4. **Careful with labels for people:** *the blind => blind people, *mentally ill => person with a psychological or emotional disability, Blacks / African-Americans is acceptable etc., avoid gendered usage (manmade => synthetic)
 5. **Avoid stereotypes:** don't stereotype groups of people with regard to their contribution to society, generosity, honesty, quality of culture etc.; mix depictions in traditional and non-traditional roles
 6. **Represent diversity:** show various ethnic groups in items
- These guidelines are not usually interpreted rigidly: where it can be shown that certain content is relevant to the construct under investigation, it is usually acceptable
 - Even though these procedures are implemented as a matter of course, it is very difficult to predict which items will show DIF for which groups, and items are continually reviewed for DIF

Future Research

- We need a better understanding of what leads to DIF: what test taker background variables interact with test item in what way? Right now that's mostly speculation
- Verbal protocols could help, as could well-designed experiments
- We need more DIF research in language testing as a whole, and particularly well-done DIF research (using MH, IRT)
- We need DIF detection models that work well with small samples

References

- Angoff, W.H. (1982). **Use of difficulty and discrimination indices for detecting item bias.** In R.A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore: Johns Hopkins University Press.
- Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland, & H. Wainer (Eds.), **Differential item functioning** (p. 3-24). Hillsdale, NJ: Lawrence Earlbaum.
- Bachman, L.F., Davidson, F., Ryan, K., Choi, I.C. (1993). **An investigation into the comparability of two tests of English as a foreign language: the Cambridge-TOEFL comparability study.** Cambridge: CUP.
- Baker, F.B. (1981). A criticism of Scheuneman's item bias technique. **Journal of Educational Measurement**, **18**, 59-62.

- Bond, L. (1993). Comments on the O'Neill & McPeck paper. In P.W. Holland, & H. Wainer (Eds.), **Differential item functioning** (p. 277-280). Hillsdale, NJ: Lawrence Erlbaum.
- Bridgeman, B. & Schmitt, A. (1997). Fairness issues in test development and administration. In W.W. Willingham, & N.S. Cole (Eds.), **Gender and fair assessment** (pp. 185-226). Mahwah, NJ: Lawrence Erlbaum.
- Camilli, G., & Shepard, L.A. (1994). **Methods for identifying biased test items**. Thousand Oaks: Sage.
- Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. **Language Testing**, 2, 2, 155-163.
- Dorans, N.J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. **Applied Measurement in Education**, 2, 217-233.
- Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland, & H. Wainer (Eds.), **Differential item functioning** (p. 35-66). Hillsdale, NJ: Lawrence Erlbaum.
- Eels, K., Havighurst, R.J., Herrick, V.E., & Tyler, R.W. (1951). **Intelligence and cultural differences**. Chicago: University of Chicago Press.
- Elder, C. (1996). The effect of language background on "foreign" language test performance: The case of Chinese, Italian, and Modern Greek. **Language Learning**, 46, 2, 233-282.
- ETS (2003). **Fairness review guidelines**. Princeton, NJ: Author. Available: http://www.ets.org/Media/About_ETS/pdf/overview.pdf
- ETS (2004). **ETS international principles for fairness review of assessments**. Princeton, NJ: Author. Available: http://www.ets.org/Media/About_ETS/pdf/frintl.pdf
- Jensen, A.R. (1980). **Bias in mental testing**. New York: Free Press.
- Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. **Language Testing**, 18, 1, 89-114.
- Kunnan, A.J. (1990). DIF in native language and gender groups in an ESL placement test. **TESOL Quarterly**, 24, 741-746.
- Kunnan, A.J. (2000). Fairness and justice for all. In A.J. Kunnan (Ed.), **Fairness and validation in language assessment** (pp. 1-14). Cambridge: CUP.
- Lee, Y.W., Breland, H., Muraki, E. (2004). **Comparability of TOEFL CBT writing prompts for different native language groups** (TOEFL RR-77). Princeton, NJ: ETS. Available: <http://www.ets.org/Media/Research/pdf/RR-04-24.pdf>
- Muniz, J., Hambleton, R.K., Xing, D. (2001). Small sample studies to detect flaws in item translations. **International Journal of Testing**, 1, 2, 115-135.

- O'Neill, K.A. & McPeck, W.M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland, & H. Wainer (Eds.), **Differential item functioning** (p. 255-276). Hillsdale, NJ: Lawrence Earlbaum.
- Penfield, R.D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. **Applied Measurement in Education, 14**, 3, 235-259.
- Ramsey, P.A. (1993). Sensitivity review: The ETS experience as a case study. In P.W. Holland, & H. Wainer (Eds.), **Differential item functioning** (p. 367-389). Hillsdale, NJ: Lawrence Earlbaum.
- Ravitch, D. (2003). **The language police**. New York: Alfred Knopf.
- Rooney, J.P. (1987). A response from Golden Rule to "ETS on 'Golden Rule.'" **Educational Measurement: Issues and Practice, 6**, 5-8.
- Ryan, K., & Bachman, L.F. (1992). Differential item functioning on two tests of EFL proficiency. **Language Testing, 9**, 1, 12-29.
- Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. **Language Testing, 8**, 2, 95-111.
- Scheunemann, J.D. (1979). A method of assessing bias in test items. **Journal of Educational Measurement, 16**, 143-152.
- Shepard, L.A. (1981). Identifying bias in test items. In B.F. Green (Ed.), **New directions in testing and measurement: Issues in testing-Coaching, disclosure, and test bias** (pp. 79-104). San Francisco: Jossey-Bass.
- Shohamy, E. (2000). Fairness in language testing. In A.J. Kunnan (Ed.), **Fairness and validation in language assessment** (pp. 15-19). Cambridge: CUP.
- Takala, S., & Kaftandjieva, F. (2000). Test fairness: a DIF analysis of an L2 vocabulary test. **Language Testing, 17**, 3, 323-340.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland, & H. Wainer (Eds.), **Differential item functioning** (p. 67-113).
- Zumbo, B.D. (1999). **A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores**. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense. Available: <http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html>.