Consequences of Testing

PSY 600K
Dr. Alyssa Gibbons
April 12, 2018

Agenda

- Consequences of testing.
  - LOTS of things we could consider here – many are hard to test formally.
- One we can and do often test is **bias.**
  - Defining bias:
    - Measurement bias
    - Predictive bias
- This will set us up for **measurement equivalence** next week.

Consequences of Testing

- Why is this relevant to the validity of a test?
  - If validity is about the **inferences** we want to make from our test scores… thinking about the consequences or implications of those inferences is part of determining whether those inferences are appropriate or justifiable.
- Includes consequences to individuals and to society!
- Partially judgment-based… but there is empirical evidence we can gather about some types of consequences. We'll focus on two today:
  - Bias
  - Precision of individual test scores.

Fairness and Bias

- From the SIOP *Principles*:
  - "**Fairness** is a social rather than a psychometric concept. Its definition depends on what one considers to be fair."
    - Equal outcomes? Equal treatment? Equal opportunity to learn? Equal prediction?
  - "**Bias** refers to any construct-irrelevant source of variance that results in systematically higher or lower scores for identifiable groups of examinees."
    - Keys: **construct-irrelevant, systematically, identifiable groups.**

Two Kinds of Bias

- **Measurement** bias: members of different (identifiable) groups get systematically higher or lower scores due to **construct-irrelevant** factors.
- **Predictive** bias: the relationship (regression line) between predictor and criterion differs between subgroups.

Measurement Bias

- Most research on this question is in the context of majority-minority group differences on selection tests.
  - Can be applied any time you are concerned about potential differences in consequences for relevant groups.
- It's taken us a long time to figure out how to assess and address this.
  - Understanding the history is actually quite important.

A Brief History of Measurement Bias

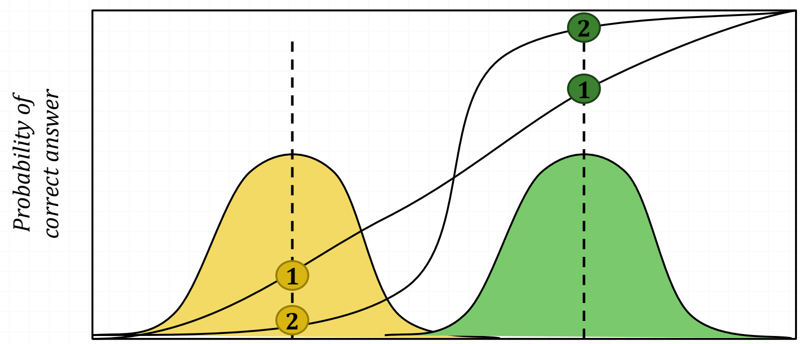
- **Approach #1: Equal Discrimination**
  - Fair items do not measure better for one group than for another.
  - Implies **equal item discrimination parameters** – point-biserial or biserial correlations.
- Problematic because:
  - It rarely happens that way!
  - Relies on comparing (sig testing differences in) correlations.
    - Sensitive to sample size, also to restriction of range.
    - So differences between groups in total test variance or item variance can have very large effects on these correlations!
    - Very difficult to be confident in our conclusions in this approach.

A Brief History of Measurement Bias

- **Approach #2: Relative Item Difficulties**
  - Even if groups have different mean ability levels, a hard item for Group 1 should still be a hard item for Group B.
  - Calculate the item difficulties for both groups and rank-order them. If the rank orders are not the same, item bias is present.
- Problematic because:
  - **Lord's paradox:** if two items have different discrimination parameters **and** there is a mean difference in ability between the groups, this can result in opposite rank-orders for the two groups **even if** the item functions identically across groups.
    - Item functioning here is in the IRT sense.
  - So the logic here is flawed.

Lord's Paradox

- Two items with different discrimination parameters.
- Two groups with different ability distributions.
- There is **no** differential functioning here!



A Brief History of Measurement Bias

- **Approach #3: Factor Analysis**
  - Items should load on the same factor to the same extent in all groups.
  - Implies equivalent (or highly similar) factor loadings.
  - Multiple Group Confirmatory Factor Analysis
    - More on this next time!
- **Approach #4: Chi-Square Indices**
  - Individuals with the same total score (proxy for true score) should have the same probability of getting an item right.

Chi-Square Indices

- Divide each group into 5 or so "bins" based on total score.
- Calculate probability of getting the item right in each bin.
- Use chi-square to compare whether probability distribution is equal for both groups.

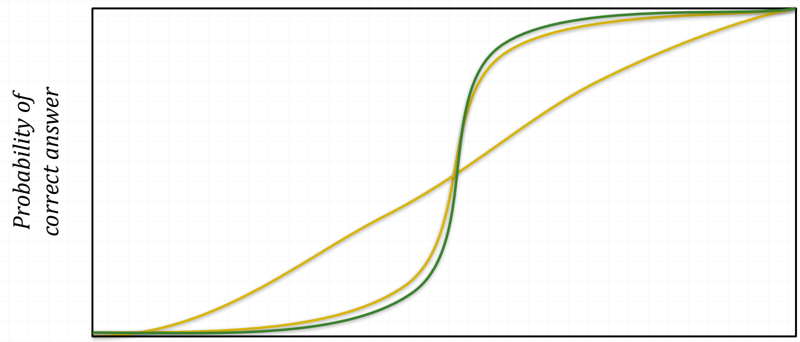
	< 25	25-39	40-59	60-74	➢ 75
Group A	.10	.20	.40	.20	.10
Group B	.25	.30	.25	.15	.5

- Problematic because:
  - We're **trying** to get a nonsignificant chi-square (prove the null hypothesis).
  - Very sensitive to sample size + arbitrary choices.

Current Approach to Measurement Bias

- In IRT: **differential item functioning**.
  - Are the item response curves (essentially) the same for both groups?
  - IRT allows us to estimate the latent trait directly rather than using total test score as a proxy.
    - More on this later!
- This means we can more directly evaluate whether people who really do have the same **true** score have the same probability of getting an item right.
  - Regardless of group trait distributions.
- This gets at a fundamental issue of fairness: do people with the same level of the trait get the same score?

Differential Item Functioning



A Brief History of Predictive Bias

- A similar story… many attempts and failures…
- **Approach #1: Single-Group Validity**
  - Concern: can a test be valid for one group and not valid for another?
  - $r(A) > 0$, $r(B) = 0$
  - Again, showing this requires proving a null hypothesis ($r(B) = 0$), heavy reliance on significance testing.
  - Reviews of these studies suggest that the proportion of studies finding single-group validity is… just about the proportion you'd expect by chance.
  - Pretty thoroughly discredited on multiple grounds.

A Brief History of Predictive Bias

- **Approach #2: Differential Validity**
  - Bias occurs when validity coefficients are not (essentially) equal for both groups.
  - $r(A) > r(B)$
  - Originally: significance tests on the difference between correlations.
    - You know how I feel about this!
  - Drasgow (1982) simulation study: Even a **very** biased test produced a difference in correlations of about .01.
    - But it did make a **substantial** difference in who was selected.
    - Not a very powerful tool for detecting bias!

A Brief History of Predictive Bias

- **Approach #3: Differential Prediction**
  - Comparing the regression relationships between predictor and criterion for both groups (a little different from the zero-order correlations).
    - Moderated multiple regression, with group as the moderator variable.
  - Either slope or intercept differences indicate bias.
- SIOP *Principles*:
  - "For White-African American and White-Hispanic comparisons, slope differences are rarely found; while intercept differences are not uncommon, they typically take the form of overprediction of minority group performance."
  - However…

Dragsow & Kang (1984)

- Simulation comparing power for differential validity & differential prediction under various conditions.
- "Our results show that the differential validity analysis should *never* be used to study measurement equivalence."
- Differential prediction is more powerful…
  - … under fairly specific circumstances.
- If you have enough power for differential prediction, you probably have a large enough sample size for DIF analysis.
  - Can identify individual problem items and remove them – don't have to throw out the whole test!

More Issues in Differential Prediction

- Requires an unbiased criterion!
- Group mean differences do not necessarily imply differential prediction… or vice versa.
  - These are truly distinct types of bias.
  - Valuable to test both… or at least to test for differential prediction even if there are not mean differences.
- If you do have differential prediction, can you just use a separate regression line for each group?
  - Select people based on their predicted criterion scores rather than their observed test scores?
  - Highly unpopular and legally questionable!
- Reliability of the test also affects power.
  - Improving reliability makes it more likely that you can detect differential prediction… so you can make your test less fair by making it more reliable !

Questions?

For next time: Measurement Equivalence
Skim: Vandenberg & Lance (2000);
also Mplus manual pp. 421-435

Lab Friday: Project Work Day!