The Use of Test Scores for Secondary Analysis: An Economist's Perspective

Jesse Rothstein

University of California, Berkeley

June 2019

Outline

Psychometrics and Economics

- Psychometricians and econometricians don't seem to interact much.

- This has costs:
  - Uses of measures should align with the way the measures are made, and often they don't.
  - Measurement construction could profitably take more account of eventual uses.

The kinds of analyses secondary users want to do

- Focus on two simple linear regressions:

  Dependent variable $\quad \theta_i = X_i\beta + \epsilon_i$
  Independent variable $\quad Y_i = \theta_i\gamma + X_i\delta + u_i$

  - θ_i is the individual's "true" achievement; we have only a measure $\hat{\theta}_i$.
  - X_i is a set of individual characteristics and/or policy variables
  - Y_i is some outcome – e.g., wages.

The kinds of analyses secondary users want to do

- Focus on two simple linear regressions:

  Dependent variable $\quad \theta_i = X_i\beta + \epsilon_i$
  Independent variable $\quad Y_i = \theta_i\gamma + X_i\delta + u_i$

  - θ_i is the individual's "true" achievement; we have only a measure $\hat{\theta}_i$.
  - X_i is a set of individual characteristics and/or policy variables
  - Y_i is some outcome – e.g., wages.

- Set aside questions of causality and asymptotics – focus on large-sample estimation of linear regressions, using $\hat{\theta}_i$ in place of θ_i.

- What properties do we need $\hat{\theta}_i$ to have? What properties does it have?

Issues that users need to understand

- Focus on what Braun & von Davier (2017) call "Large Scale Assessment Surveys" (LSAS – e.g., NAEP, TIMSS, PISA)
  - Designed primarily to provide group-level score distributions, at the level of the country, state, or demographic group.
  - Tests are short, with multiple test forms.
  - Much effort goes into fixing the domain & coverage.
  - Individual proficiency estimates are a side effect, not the goal.
- Two types of issues:
  1. Scaling is arbitrary.
  2. Individual proficiency measures don't fit economists' mental categories – they aren't unbiased estimates with classical measurement error.

Outline

Scaling: Achievement does not have an interval scale

- Achievement is ordinal, not cardinal
  - A test can say $\theta_{\text{brian}} > \theta_{\text{jesse}}$, but the magnitude of $\theta_{\text{brian}} - \theta_{\text{jesse}}$ is indeterminate – not just not identified, but not well defined.
  - Any statement about θ that is not also true of θ^2, $\sqrt{\theta}$, $\ln\theta$, $\exp\theta$, $\mathbf{1}(\theta > c)$, or any other (weakly) monotonic transformation is a claim both about true achievement and about the chosen scale.
- This is a problem for linear regression!

Scaling: Achievement does not have an interval scale

- Achievement is ordinal, not cardinal
  - A test can say $\theta_{\text{brian}} > \theta_{\text{jesse}}$, but the magnitude of $\theta_{\text{brian}} - \theta_{\text{jesse}}$ is indeterminate – not just not identified, but not well defined.
  - Any statement about θ that is not also true of θ^2, $\sqrt{\theta}$, $\ln\theta$, $\exp\theta$, $\mathbf{1}(\theta > c)$, or any other (weakly) monotonic transformation is a claim both about true achievement and about the chosen scale.
- This is a problem for linear regression!
- Options:
  - ~~Rely on Z-scores to solve the problem.~~

Scaling: Achievement does not have an interval scale

- Achievement is ordinal, not cardinal
  - A test can say $\theta_{\text{brian}} > \theta_{\text{jesse}}$, but the magnitude of $\theta_{\text{brian}} - \theta_{\text{jesse}}$ is indeterminate – not just not identified, but not well defined.
  - Any statement about θ that is not also true of θ^2, $\sqrt{\theta}$, $\ln\theta$, $\exp\theta$, $\mathbf{1}(\theta > c)$, or any other (weakly) monotonic transformation is a claim both about true achievement and about the chosen scale.
- This is a problem for linear regression!
- Options:
  - ~~Rely on Z-scores to solve the problem.~~
  - ~~Rely on psychometrics / item response theory (IRT) to solve the problem.~~

Scaling: Achievement does not have an interval scale

- Achievement is ordinal, not cardinal
  - A test can say $\theta_{\text{brian}} > \theta_{\text{jesse}}$, but the magnitude of $\theta_{\text{brian}} - \theta_{\text{jesse}}$ is indeterminate – not just not identified, but not well defined.
  - Any statement about θ that is not also true of θ^2, $\sqrt{\theta}$, $\ln\theta$, $\exp\theta$, $\mathbf{1}(\theta > c)$, or any other (weakly) monotonic transformation is a claim both about true achievement and about the chosen scale.
- This is a problem for linear regression!
- Options:
  - ~~Rely on Z-scores to solve the problem.~~
  - ~~Rely on psychometrics / item response theory (IRT) to solve the problem.~~
  - Check robustness to transformations (though typically need to bound the space).
  - Scale to an external interval metric.
- Tim Bond is the expert – see his talk!

IRT doesn't solve the scaling problem

- Many tests are scaled using Item Response Theory (IRT)
  - Let $r_{ij} = 1$ if student i gets item j right.
  - An IRT model specifies $\Pr\{r_{ij} = 1 \mid \theta_i, \psi_j\}$.
  - The "3 parameter logistic" (3PL) IRT model:

$$\Pr\left\{r_{ij} = 1 \mid \theta_i, \psi_j = \{a_j, b_j, c_j\}\right\} = c_j + (1 - c_j)\,\frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}}$$

  b_j is difficulty, a_j is discrimination, and c_j is guessability.

IRT doesn't solve the scaling problem

- Many tests are scaled using Item Response Theory (IRT)
  - Let $r_{ij} = 1$ if student i gets item j right.
  - An IRT model specifies $\Pr\{r_{ij} = 1 \mid \theta_i, \psi_j\}$.
  - The "3 parameter logistic" (3PL) IRT model:

$$\Pr\left\{r_{ij} = 1 \mid \theta_i, \psi_j = \{a_j, b_j, c_j\}\right\} = c_j + (1 - c_j)\,\frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}}$$

    b_j is difficulty, a_j is discrimination, and c_j is guessability.

- This defines a scale. But observed responses are equally compatible with any other scale $\tilde{\theta} = g(\theta)$ for $g(\cdot)$ (strictly) monotonic:

$$\Pr\left\{r_{ij} = 1 \mid \tilde{\theta}_i, \psi_j\right\} = c_j + (1 - c_j)\,\frac{e^{a_j\left(g^{-1}(\tilde{\theta}_i) - b_j\right)}}{1 + e^{a_j\left(g^{-1}(\tilde{\theta}_i) - b_j\right)}}$$

Outline

Proficiency measure: Goals and fit

- Tests are short, so θ_i not identified.
- Goals as analysts:
  1. Characterization of f(θ) for pre-specified groups (nationalities, races).
  2. Use $\hat{\theta}$ as a dependent variable
  3. Use $\hat{\theta}$ as an independent variable

Proficiency measure: Goals and fit

- Tests are short, so θ_i not identified.
- Goals as analysts:
  1. Characterization of $f(\theta)$ for pre-specified groups (nationalities, races).
  2. Use $\hat{\theta}$ as a dependent variable
  3. Use $\hat{\theta}$ as an independent variable

- Many LSAS use "plausible values," random draws from the posterior distribution of θ given item responses and student background characteristics Z.
  - PVs suffice but are unnecessary for goal 1.
  - Suitability for goals 2 and 3 depends on the specific model, and on X and Z.
  - Accommodating goals 2 and 3 for all potential regression models requires end users to model item responses directly.

What is a "plausible value"?

- Assume $\theta_i \sim \mathbb{N}\left(\mu(Z_i), \sigma^2_\theta(Z_i)\right)$
- IRT model gives likelihood of observed responses $\text{IRT}(R_i \mid \theta_i)$.
- By Bayes Rule, posterior distribution of θ is

$$\begin{aligned} p(\theta \mid R, Z) &= \frac{p(R \mid \theta, Z)p(\theta \mid Z)}{p(R \mid Z)} \\ &= \frac{\text{IRT}(R \mid \theta)\phi(\theta; \mu(Z), \sigma^2_\theta(Z))}{\int \text{IRT}(R \mid t)\phi(t; \mu(Z), \sigma^2_\theta(Z))dt} \end{aligned}$$

What is a "plausible value"?

- Assume $\theta_i \sim \mathbb{N}\left(\mu(Z_i), \sigma^2_\theta(Z_i)\right)$
- IRT model gives likelihood of observed responses $\text{IRT}(R_i \mid \theta_i)$.
- By Bayes Rule, posterior distribution of θ is

$$\begin{aligned} p(\theta \mid R, Z) &= \frac{p(R \mid \theta, Z)p(\theta \mid Z)}{p(R \mid Z)} \\ &= \frac{\text{IRT}(R \mid \theta)\phi(\theta; \mu(Z), \sigma^2_\theta(Z))}{\int \text{IRT}(R \mid t)\phi(t; \mu(Z), \sigma^2_\theta(Z))dt} \end{aligned}$$

- Step 1: Estimate $\mu(Z_i), \sigma^2_\theta(Z_i)$.
- Step 2: Take K draws from posterior distributions of $\hat{\mu}$ and $\hat{\sigma}^2_\theta$, then from $p(\theta \mid R, Z)$ given these.

What is a "plausible value"?

- Assume $\theta_i \sim \mathbb{N}\left(\mu(Z_i), \sigma^2_\theta(Z_i)\right)$
- IRT model gives likelihood of observed responses $\text{IRT}(R_i \mid \theta_i)$.
- By Bayes Rule, posterior distribution of θ is

$$\begin{aligned} p(\theta \mid R, Z) &= \frac{p(R \mid \theta, Z)p(\theta \mid Z)}{p(R \mid Z)} \\ &= \frac{\text{IRT}(R \mid \theta)\phi(\theta; \mu(Z), \sigma^2_\theta(Z))}{\int \text{IRT}(R \mid t)\phi(t; \mu(Z), \sigma^2_\theta(Z))dt} \end{aligned}$$

- Step 1: Estimate $\mu(Z_i), \sigma^2_\theta(Z_i)$.
- Step 2: Take K draws from posterior distributions of $\hat{\mu}$ and $\hat{\sigma}^2_\theta$, then from $p(\theta \mid R, Z)$ given these.
- Analogies: Empirical Bayes (for posterior mean), multiple imputation

Comparing means to ends 1: Characterizing p(θ | G) for pre-specified groups

- Suppose we want to estimate only $\mathbb{E}[\theta \mid G]$ and $V(\theta \mid G)$.
- A noisy but unbiased estimate would identify $\mathbb{E}[\theta \mid G]$ but overstate $V(\theta \mid G)$.

Comparing means to ends 1: Characterizing p(θ | G) for pre-specified groups

- Suppose we want to estimate only $\mathbb{E}[\theta \mid \mathrm{G}]$ and $\mathrm{V}(\theta \mid \mathrm{G})$.
- A noisy but unbiased estimate would identify $\mathbb{E}[\theta \mid \mathrm{G}]$ but overstate $\mathrm{V}(\theta \mid \mathrm{G})$.
- PVs avoid bias, but add unnecessary steps.

Comparing means to ends 1: Characterizing $p(\theta \mid G)$ for pre-specified groups

- Suppose we want to estimate only $\mathbb{E}[\theta \mid G]$ and $V(\theta \mid G)$.
- A noisy but unbiased estimate would identify $\mathbb{E}[\theta \mid G]$ but overstate $V(\theta \mid G)$.
- PVs avoid bias, but add unnecessary steps.
  - The estimates $\mu(Z_i), \sigma^2_\theta(Z_i)$ are sufficient for the goal, if $G \subseteq Z$.

$$p(\theta \mid G) = \int p(\theta \mid Z) df(Z \mid G) = \int \phi(\theta; \mu(Z), \sigma^2_\theta(Z)) df(Z \mid G)$$

  - Possible efficiency gains (akin to poststratification) from integrating over $p(Z \mid G)$.
  - PVs don't add anything once we have $\mu(Z), \sigma^2_\theta(Z)$.

Comparing means to ends 1: Characterizing $p(\theta \mid G)$ for pre-specified groups

- Suppose we want to estimate only $\mathbb{E}[\theta \mid G]$ and $V(\theta \mid G)$.
- A noisy but unbiased estimate would identify $\mathbb{E}[\theta \mid G]$ but overstate $V(\theta \mid G)$.
- PVs avoid bias, but add unnecessary steps.
  - The estimates $\mu(Z_i), \sigma^2_\theta(Z_i)$ are sufficient for the goal, if $G \subseteq Z$.

$$p(\theta \mid G) = \int p(\theta \mid Z) df(Z \mid G) = \int \phi(\theta; \mu(Z), \sigma^2_\theta(Z)) df(Z \mid G)$$

  - Possible efficiency gains (akin to poststratification) from integrating over $p(Z \mid G)$.
  - PVs don't add anything once we have $\mu(Z), \sigma^2_\theta(Z)$.
- We might also want nonparametric estimate of $p(\theta \mid G)$, but PVs rely on parametric $p(\theta \mid Z)$.

Comparing means to ends 2: Use θ as a dependent variable

- A noisy but unbiased estimate would be no problem here.
- PVs can work, sometimes.
  - PV has two components, $\hat{\theta}_{ik}^{PV} = \bar{\theta}_{i}^{EAP} + u_{ik}$
    - $\bar{\theta}_{i}^{EAP} \approx \mathbb{E}[\theta \mid R_i, Z_i]$ is the posterior mean.
    - u_{ik} is a generated random number.
  - Bias of regression of $\hat{\theta}_{ik}^{PV}$ on X_i is the same as with $\bar{\theta}_{i}^{EAP}$
    - Unbiased if $X \subseteq Z$ (by iterated expectations); biased otherwise.

Comparing means to ends 2: Use θ as a dependent variable

- A noisy but unbiased estimate would be no problem here.
- PVs can work, sometimes.
  - PV has two components, $\hat{\theta}_{ik}^{PV} = \bar{\theta}_{i}^{EAP} + u_{ik}$
    - $\bar{\theta}_{i}^{EAP} \approx \mathbb{E}[\theta \mid R_i, Z_i]$ is the posterior mean.
    - u_{ik} is a generated random number.
  - Bias of regression of $\hat{\theta}_{ik}^{PV}$ on X_i is the same as with $\bar{\theta}_{i}^{EAP}$
    - Unbiased if $X \subseteq Z$ (by iterated expectations); biased otherwise.
  - Variance. Between-PV variation reflects two components
    1. Estimation error in parameters of $\mu(Z)$
    2. Random draws from the distribution around estimated $\mu(Z)$.

    - #2 reduces efficiency.
    - #1 is important (e.g., consider $X = Z$ case).

Comparing means to ends 3: Use θ as an independent variable

- $Y = \theta\gamma + X\delta + u$
- A noisy but unbiased estimate $\hat{\theta}$ would attenuate γ and bias δ.
- PVs can work, if regression is "congenial" with conditioning model $p(\theta \mid Z)$.
  - Roughly, PVs work if they recover joint distribution of $\{X, Y, \theta\}$.
  - If $\{X, Y\} \subseteq Z$, can estimate exactly one model $p(Y \mid \theta, X)$, but not necessarily the one we want (Schofield et al., 2015).
- Not a lot of good options here.
  - With item response data, MESE model (Schofield 2012); see below.
  - Unbiased estimate with known reliability, and EIV correction.
  - Instrumental variables with two sub-test scores (unbiased estimates)

What to do about it?

- Additional reported statistics (esp. unconditionally unbiased estimates, posterior means, and specific conditioning variables Z) can help.
- Other solutions involve releasing item responses, and relying on researchers to build models for them.
- Need more sophistication from researchers, as well as more support from testmakers.

Marginal Maximum Likelihood estimator - dependent variable case

How to solve the dependent variable problem with item-level data.

- Model:
  1. Research model: $\theta = X\beta + \epsilon$, $\epsilon \sim \mathbb{N}(0, \sigma^2)$ yields $p(\theta \mid X; \beta, \sigma)$.
  2. IRT model: $IRT(R_i \mid \theta_i; \psi)$
- Observed data likelihood:

$$p(R_i|X_i) = \int p(R_i \mid \theta, X_i) dp(\theta \mid X_i; \beta, \sigma)$$
$$= \int IRT(R_i \mid \theta; \psi) dp(\theta \mid X_i; \beta, \sigma)$$

- Solve by numerical integration, and maximize over $\{\beta, \sigma \psi\}$.

Mixed Equations Structural Estimator - independent variable case

How to solve the independent variable problem with item-level data (Schofield et al. 2012)

- Model:
  - Research model: $Y = \theta\gamma + X\delta + u$, $u \sim \mathbb{N}(0, \sigma^2)$ yields $p_{RM}(Y \mid \theta, X; \gamma, \delta, \sigma)$.
  - IRT model: $IRT(R_i \mid \theta_i, \psi)$
  - Custom conditioning model: $\theta \mid X \sim \mathbb{N}(X\pi, \tau^2)$ yields $p_{CCM}(\theta \mid X_i)$.
- Observed data likelihood:

$$
\begin{aligned}
p(Y_i, R_i \mid X_i) &= \int p(Y_i, R_i, \theta \mid X_i) p(\theta \mid X_i) d\theta \\
&= \int p(Y_i \mid R_i, \theta, X_i) p(R_i \mid \theta, X_i) p(\theta \mid X_i) d\theta \\
&= \int p_{RM}(Y_i \mid \theta, X_i) IRT(R_i \mid \theta, \psi) p_{CCM}(\theta \mid X_i) d\theta
\end{aligned}
$$

- Solve by numerical integration, and maximize over $\{\gamma, \delta, \sigma, \tau, \psi\}$.

Outline

Conclusion

- Economists can't take our measures for granted.
- We are used to thinking about classical measurement error.
- Plausible values are not that!
- Need to think more about aligning measures to analyses in education.
- This requires changes in secondary analysts' practice – though test makers could help too.