EPSE 594: Meta-Analysis: Quantitative Research Synthesis

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

March 31, 2022

Psychometric issues in meta-analysis and beyond

In psychometrics, we are often very concerned with issues of *measurement*, namely:

- Reliability: how variable, or *imprecise*, a measurement process is.
- Validity: how well (how *accurately* and *precisely*) the measurement captures the phenomenon it is trying to quantify.

A concept that is often hidden or implied is:

- Calibration: how *accurate* (usually on average) a measurement process is for what its trying to measure (see, e.g., Kroc & Zumbo 2018).

With this terminology in mind, one can think (broadly speaking):

$$\text{reliability} + \text{calibration} = \text{validity}$$

In this way, measurement quality is analogous to estimator quality:

$$\text{variance} + \text{bias} = \text{quality}$$

Psychometric issues in meta-analysis and beyond

Classically, one proposes the following framework:

- Each subject (e.g. person) has a unique *true value (score)*, T, of some particular phenomenon of interest.
- This true value cannot be measured directly; instead, we observe (measure) only a proxy for it; this is the *observed score*, X.
- This observed score may differ from the true score; thus we propose a generic *measurement error model*:

$$X = T + E,$$

  where E denotes the *measurement error*.

- Usually, further assumptions are then imposed on the structure of the errors to more accurately model a real-life phenomenon and measurement process.

Psychometric issues in meta-analysis and beyond

- The *classical test theory* (CTT) proposes that the observed scores *balance* on the true score for every individual over repeated, memoryless reapplications of the same measurement.
- Formally, CTT posits:

$$X = T + E,$$

where $\mathbb{E}(X \mid \text{individual } i) = T$.

- Equivalently,

$$X = T + E,$$

where $\mathbb{E}(E \mid \text{individual } i) = 0$ and the true score must be fixed for every individual (see Kroc & Zumbo 2020 for details).

- This is the standard *measurement error model* employed in virtually all of the behavioural sciences.
- But just how reasonable is this model?

Psychometric issues in meta-analysis and beyond

- Two important implications of the CTT are:
  - $\mathbb{E}(E \mid \text{individual } i) = 0$; i.e. the errors will *balance* out to zero for every individual (calibration).
  - $\text{Cov}(E, T) = 0$; i.e., the errors are uncorrelated with the true scores.
- A stronger version of CTT would presume that E and T are actually *independent*. This is implied if one assumes everything is normally distributed, but that is *not* part of the classical test theory.
- If one assumes $E \perp T$, then this is commonly referred to as a *strong errors-in-variables* or *errors independent of true scores* measurement error model, common in economics.

Psychometric issues in meta-analysis and beyond

- Many other types of measurement error models exist.
- *Classical (Spearman-Pearson) measurement error*, common in the health sciences, proposes $X = T + E$ where $\mathbb{E}(E \mid T) = 0$.
  - So errors only need balance over individuals with the same true score, but not necessarily on each individual.
- *(Weak) errors-in-variables* models, common in economics, assume $X = T + E$ where $\mathbb{E}(E) = 0$ and $\text{Cov}(T, E) = 0$.
  - This looks almost like the CTT, but crucially, the errors only have to balance *over the entire population/sample*, not on each individual within the population/sample, but they cannot correlate with the true scores.
- *Berkson measurement error*, common in biochemistry and epidemiology, proposes $X = T + E$ where $\mathbb{E}(E \mid X) = 0$
  - So errors balance over observed scores, but not necessarily over individuals or over true scores.

Psychometric issues in meta-analysis and beyond

- Other measurement error models/frameworks in use:
  - *Weakly calibrated error models*: $X = T + E$ where only $\mathbb{E}(E) = 0$ is assumed. So errors can correlate with true or observed scores, but must balance out over the entire population/sample.
  - *Item response theory*: can't quite be written as $X = T + E$, but has a similar flavour as the CTT model.
  - *Generalizability theory*: An extension of CTT to multiple sources of measurement error.
  - *Generalized measurement error modelling*: An extension of all of the above to *random-variable-valued measurements* (RVVMs), useful when a single number/score cannot be assigned to a sample measurement/observation.
- See Kroc & Zumbo (2020) and Kroc (2020) for details.

Psychometric issues in meta-analysis

- In the context of meta-analysis in the behavioural and health sciences, one most commonly encounters quantities with measurement error described by the CTT.
- Specifically, one may want to synthesize effects based on *scales* derived from some kind of factor analytic construction of a *latent variable*, like quantifying *well-being* with one or several different subscales.
- In this case, the effect sizes of interest do not come from directly observable variables. That is, we can't explicitly measure "well-being" in an experimental and a control group, so instead we quantify "well-being" by some kind of survey and then create one or more latent variables out of those survey responses that capture most of the variation in the survey responses.

Psychometric issues in meta-analysis

- All such analyses should come equipped with some empirical measure of *reliability* of the (sub)scales, most often quantified by a Cronbach's α (though there are MANY other quantifiers).
- There may also be other measures of how *reliable* or *valid* the measurements (observed) are for the actual phenomenon (true, unobserved) they are trying to quantify.
- Note: *this is not the same thing as sampling error*.
- Sampling error occurs because our sample will not capture every relevant feature of the overall population. It quantifies how much we would expect our estimates to change upon repeated resampling from the population.

Psychometric issues in meta-analysis

- In contrast, *measurement error* speaks to how well our measurement process captures the phenomenon it is trying to quantify.
- The CTT notion of measurement error imagines readministering the survey/test to each sample individual (assuming we wipe their memories of the previous survey/test clean), so then measurement error quantifies how much we would expect their survey/test responses to change upon this readministration.
- Hypothetically, one could have data and estimators that are subject to:
  - Only sampling error, no measurement error (the most classical type of assumed data).
  - Only measurement error, no sampling error (could only occur with census-level data).
  - Both sampling and measurement error (the norm for any data context).

Psychometric issues in meta-analysis

In the context of meta-analysis, one may want study weights to explicitly account for the *reliability* or *validity* of a measurement from a particular study.

- In practice, you really only see this done with estimates of *reliability*.
- For simple measurement error situations, extra variance due to measurement error (i.e. imperfect reliability) will have an *attenuating* effect on model estimates; i.e. measurement error tends to cause our estimates to *shrink towards the null*.
- However, if we could adjust our estimates before meta-analyzing them, then we could potentially remove (at least some) of this *attenuation bias*.
- Warning: Attenuation towards the null due to measurement error is only mathematically guaranteed to occur if a *single* variable is subject to measurement error (without range restriction).

Psychometric issues in meta-analysis

Recall:

- *Reliability* of a measurement X for a true score T is defined as

$$Rel(X, T) := \frac{\text{Var}(T)}{\text{Var}(X)} = \frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(E)}.$$

- Note: Reliability can *never* be directly estimated because it depends on the true score T; hence, one has to propose a *model* of how T relates to X to try and *quantify* reliability.

- Under the *classical test theory* measurement error model, if one has two *parallel measurements* X and X' for T, then reliability is also equal to:

$$Rel(X, T) = \rho^2_{XX'}$$

- X and X' are parallel measurements for T if their variances are equal, and their corresponding errors are uncorrelated.

Psychometric issues in meta-analysis

- It can be shown that if ρ is a correlation (effect size) between variables, with one subject to CTT measurement error, and if ρ_{adj} is the *corrected* correlation (hypothetically free of measurement error), then
$$\sqrt{Rel(X,T)} = \frac{\rho}{\rho_{adj}}.$$
- Thus, to adjust for attenuation due to this kind of measurement error, we need a way to *quantify reliability*.
- If you have parallel measurements, then great. But this is rare.
- So many methods have been developed to quantify reliability when only one measurement process (or non-parallel measurements) is/are available.
- Cronbach's α is the most common. Usually, all such quantifiers yield *underestimates* of the actual (theoretical) reliability.

Psychometric issues in meta-analysis

- Cronbach's α is the most common quantifier of reliability, but there are *many* (too many) others. For example:
  - Cronbach's alpha = coefficient alpha = KR 20
  - Ordinal alpha (for discrete but ordinal data)
  - Omega and ordinal omega
  - ICCs = intraclass correlation coefficients (extensions of KR 20)
  - Kendall's tau (for Spearman/nonparametric correlations, 0/1 data)
  - And many, many others.
- As a general rule, choosing between these different quantifiers is splitting hairs, like arguing whether a *p*-value of 0.04 is really that different from one that's 0.05.
- My advice: Use the quantifier that the journal/readers/field is most used to seeing. Rarely is it theoretically (and never empirically) justifiable to use one of these over the other (except in the case of nominal 0/1-style or truly discrete ordinal – NOT Likert – data).

Psychometric meta-analysis

- When people speak of a "psychometric meta-analysis," what they mean $> 99\%$ of the time is that they adjusted their effect size (usually a correlation) to account for CTT-style measurement error.
- In practice, this means replacing our theoretical equation

$$\sqrt{Rel(X, T)} = \frac{\rho}{\rho_{adj}}$$

by its sample analogue:

$$\sqrt{\alpha} = \frac{\widehat{\rho}}{\widehat{\rho}_{adj}},$$

where α is (usually) Cronbach's α.

- Thus, for any study with effect sizes subject to measurement error, we can work instead with the *adjusted effect size* (usually a correlation):

$$\widehat{\rho}_{adj,k} = \frac{\widehat{\rho}_k}{\sqrt{\alpha}}.$$

Psychometric meta-analysis

- Notice how this adjustment counteracts any attenuating effects:

$$r_{adj,k} = \frac{r_k}{\sqrt{\alpha}}.$$

- The less reliable the measurement, the smaller the α; hence, the bigger the adjustment.
- The more reliable the measurement, the closer α is to 1; hence the smaller the adjustment.
- Usually, if one is meta-analyzing studies with effect sizes that are subject to measurement error, each study will report some quantifier of reliability. Ideally, these are all the same, so that each study's adjustment is the same.
- In practice, the reported reliability quantifiers may be different. Difficult to convert from one to the other, but remember: the particular choice rarely matters much anyway.

Psychometric meta-analysis

- Adjusting effect size estimates means we also need to adjust their associated standard errors:

$$SE(r_{adj}) = \sqrt{\text{Var}(r_{adj})} = \sqrt{\frac{\text{Var}(r)}{\alpha^2}} = \frac{SE(r)}{\alpha}$$

- Now can proceed to meta-analyze as usual, but using these *adjusted* estimates of effect size and standard error.

Psychometric meta-analysis

- Psychometric meta-analysis evolved somewhat independently of mainstream meta-analysis; hence, there are some different traditions/conventions.
- The most notable of these is that usually people use a simple *sample size weighted* estimate of aggregate effect size, instead of the more typical inverse-variance weighted estimator. That is, usually one sees:

$$\overline{r} = \frac{\sum_{k=1}^{N} n_k r_k}{\sum_{k=1}^{N} n_k}.$$

- To my knowledge, there is *no* statistical justification for doing this. Plugging in the adjusted sample correlations $r_{k,adj}$ for r_k only accounts for measurement error, not population variance.
- A more statistically justified estimator (e.g., a maximum likelihood estimator) would account explicitly for sampling error via the estimated population variances and sample sizes.

Psychometric meta-analysis

- Such an MLE would look something like:

$$\bar{r} = \frac{\sum_{k=1}^{N} w_k r_k}{\sum_{k=1}^{N} w_k},$$

where

$$w_k = \frac{1}{\widehat{\sigma}_k^2 + \tau^2 + (1 - \alpha_k)}$$

for a random/mixed effects model.

- Here, the weights reflect within-study sampling uncertainty, $\widehat{\sigma}_k^2$, between-study heterogeneity, τ^2, and within-study measurement uncertainty, α_k.

- I would use an estimator like this if I were meta-analyzing effect sizes subject to measurement error, but you will rarely see this done in the literature (because of "tradition").

Psychometric meta-analysis

- Moreover, psychometric meta-analysis tends to use a different estimator for the sample variance of each study effect size:

$$\text{Var}(r_k) = \frac{(1 - \bar{r}^2)^2}{n_k - 1}.$$

- Again, there is no good statistical reason for doing this.
- The more natural estimator is

$$\text{Var}(r_k) = \frac{(1 - r_k^2)^2}{n_k - 1}.$$

- Again, this is the estimator I would use, but you will rarely see this done in the literature (because of "tradition").
- Note: Using these different estimators of sample variance will lead to different estimators of between-study heterogeneity, $\widehat{\tau}^2$.
- In practice, these differences won't probably amount to much.

Psychometric meta-analysis: example

Table 38.1 Fictional data for psychometric meta-analysis.

Study	*n*	*r*	*Criterion reliability*
University 1	130	0.24	0.75
University 2	90	0.11	0.75
Private 1	30	0.05	0.60
Private 2	25	0.17	0.60
Volunteer 1	50	0.38	0.90
Volunteer 2	65	0.50	0.90

- From Borenstein (Chapter 38), we have 6 studies assessing the validity of pre-hire work sample test to predict job performance of dental hygienists six months after hire.
- Job performance is measured using different rating scales: one for the universities, one for the private practices, and one for the nonprofits.

Psychometric meta-analysis: example

- We can calculate the adjusted effect sizes (and corresponding variances) by using the adjustment formulas from before, e.g.,

$$r_{adj} = \frac{r}{\sqrt{\alpha}}.$$

- Here, we convert

$$r_1 = 0.24,\ r_2 = 0.11,\ r_3 = 0.05,\ r_4 = 0.17,\ r_5 = 0.38,\ r_6 = 0.50$$

into

$$r_{1,a} = 0.28,\ r_{2,a} = 0.13,\ r_{3,a} = 0.06,\ r_{4,a} = 0.22,\ r_{5,a} = 0.40,\ r_{6,a} = 0.53$$

- Of course, there is an R package that will do this for you automatically, 'psychmeta'. All you have to do is feed in the raw effect sizes, standard errors or sample sizes, and the reliability quantifiers.

Psychometric meta-analysis: example

- The estimated aggregate effect after adjustment would be $\bar{r} = 0.29$ with corresponding estimated SE $= 0.14$.
- Contrast this with what we would get if we meta-analyzed the unadjusted effect sizes: $\bar{r}_{unadj} = 0.25$ with corresponding estimate SE $= 0.14$.
- So not a huge difference, but this is to be expected given the relatively large Cronbach's alphas from each study.