A transdisciplinary view of measurement error models and the variations of $X = T + E$

Edward Kroc (ed.kroc@ubc.ca) & Bruno D. Zumbo

Measurement, Evaluation, and Research Methodology
Department of Educational and Counselling Psychology, and Special Education
University of British Columbia, Vancouver, BC

April 8, 2019: NCME, Toronto, ON

Measurement error models

- Many disciplines use *measurement error models* (MEM).
- The most common MEMs take the generic form

$$X = T + E, \text{ where:}$$

  - T is the true value (score) of the random variable of interest,
  - X is an observable proxy for T (observed score),
  - E is the corresponding error induced when measuring T by X.

- Always, further assumptions are then placed on the structure of these random variables, depending on the research situation.

Measurement error models

Five common MEMs in practice:

- Classical test theory (CTT)
- Errors independent of true scores (EIT)
- Classical measurement error (CME)
- Weak errors in variables (WEV)
- Berkson measurement error (BME)

. . . and one new one introduced in Kroc & Zumbo (*JMASM*, 2018):

- Weakly calibrated errors (WCE)

Measurement error models

- Different MEMs are more common in different disciplines.
- Different MEMs are more *appropriate* in different disciplines.
  - Psychometrics: CTT
  - Surveys: CTT, EIT, CME, WEV, (WCE)
  - Econometrics: EIT, CME, WEV
  - Epidemiology: CME, BME
  - Ecology: CME, BME, (WCE)
  - Chemistry & Physics: CME, BME

What about IRT?

- The model(s) of item response theory (IRT) does *not* fall into this general class of MEMs.
- That is, IRT does *not* posit:

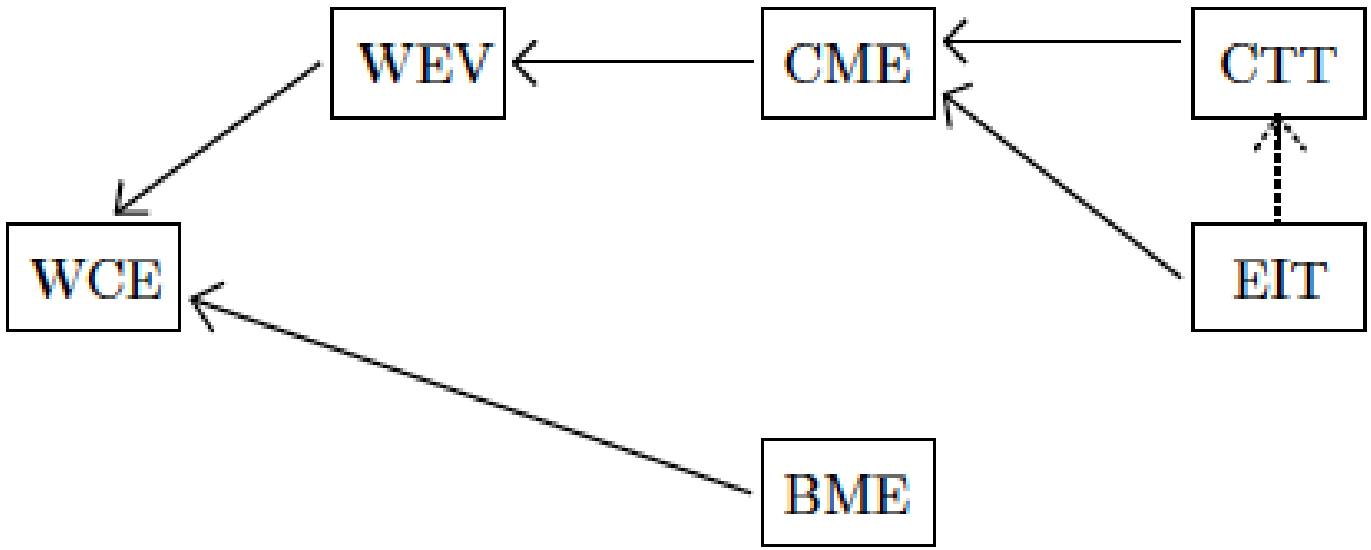
$$X = T + E$$

- Many intimate relationships between CTT and IRT; e.g. see Takane & de Leeuw, 1987; or Raykov & Marcoulides, 2016

IRT posits a measurement model, but *not* a measurement *error* model

Relationships between the MEMs

Relationships between the six MEMs:



- An arrow begins at a stronger model and points to a weaker model (i.e. one that contains the stronger model as a special case).
- Dotted arrow indicates a weaker form of inclusion.

The CTT model

- The CTT model proposes

$$X = T + E,$$

with an additional (strong) assumption that, for every individual (distinct unit in the study population), the individual's true score equals the average of their observed scores upon repeated applications of the measurement (test).

- More formally, for every individual i in the study population, there is a collection of events denoting all possible applications of the test to that individual: $\sigma(i)$. Then, we require

$$T = \mathbb{E}(X \mid \sigma(i)), \quad \text{for every } i.$$

The CTT model

- Several equivalent ways to define the CTT model (see paper).
- The exchangeability captured by the

$$T = \mathbb{E}(X \mid \sigma(i))$$

  condition is very strong: *errors balance on every individual*.
- This condition is the *defining feature* of the CTT.
- Some practitioners recognize this, while others remain confused. . .

Correct characterizations of the CTT in the literature

Raykov & Marcoulides, *Educational and Psychological Measurement*, 2016:

> *The observed score, for any prespecified individual, is a random variable pertaining to the administration of the measure to him or her. In particular, the mean of X ... [is] equal by definition to the associated true score.*

That is,

$$T = \mathbb{E}(X \mid \sigma(i)), \quad \text{for every } i.$$

Correct characterizations of the CTT in the literature

Mair, *Modern Psychometrics with R*, 2018:

> *If we present the same item i to an individual ν many times, the true score is the average of the observed scores.*

That is,

$$T = \mathbb{E}(X \mid \sigma(\nu)), \quad \text{for every } \nu.$$

Correct characterizations of the CTT in the literature

De Champlain, *Medical Education*, 2010:

> *The candidate's true score is defined as the expected value of the observed score over an infinite number of repeat administrations with the same examination.*

That is,

$$T = \mathbb{E}(X \mid \sigma(i)), \quad \text{for every } i.$$

Incorrect characterizations of the CTT in the literature

In *Educational and Psychological Measurement*, 2012:

> *[In] the classical test theory, true scores and errors are assumed to be independent.*

- This specifies the EIT model
- EIT is *not* directly comparable to CTT
- Lacks balance of errors on *every individual*
- True scores and errors need not be independent under CTT: model only implies they are uncorrelated.

CTT vs. EIT

	Test/Retest	1	2	1	2	1	2	1	2
	Individual	1	1	2	2	3	3	4	4
	T	1	1	1	1	2	2	3	3
CTT	E_1	0	0	0	0	-1	1	-1	1
	X_1	1	1	1	1	1	3	2	4
EIT	E_2	-1	-1	1	1	-1	1	-1	1
	X_2	0	0	2	2	0	4	3	3

- EIT errors do *not* balance on $i = 1$ or $i = 2$
- CTT errors are *not* independent of T

Incorrect characterizations of the CTT in the literature

In *Eurasian Journal of Educational Research*, 2018:

> *True scores and error scores are uncorrelated, there are no systematic patterns between the error scores obtained from the parallel applications of the same measurement tool, and the expected value of the error scores is zero.*

- This specifies the EIT model
- EIT is *not* directly comparable to CTT
- Lacks balance of errors on *every individual*
- True scores and errors need not be independent under CTT: model only implies they are uncorrelated.

CTT vs. EIT

	Test/Retest	1	2	1	2	1	2	1	2
	Individual	1	1	2	2	3	3	4	4
	T	1	1	1	1	2	2	3	3
CTT	E_1	0	0	0	0	-1	1	-1	1
	X_1	1	1	1	1	1	3	2	4
EIT	E_2	-1	-1	1	1	-1	1	-1	1
	X_2	0	0	2	2	0	4	3	3

- EIT errors do *not* balance on $i = 1$ or $i = 2$
- CTT errors are *not* independent of T

Incorrect characterizations of the CTT in the literature

In *Educational and Psychological Measurement*, 2018:

> *By construction, the expected value of E is 0, while the relationship $Cov(T, E) = 0$ is assumed.*

- This specifies the WEV model
- WEV is two MEMs weaker than the CTT
- Lacks balance of errors on *every individual*
- Lacks balance of errors over *individuals with the same true score*

CTT vs. WEV

	Test/Retest	1	2	1	2	1	2	1	2
	Individual	1	1	2	2	3	3	4	4
	T	1	1	1	1	2	2	3	3
CTT	E_1	0	0	0	0	-1	1	-1	1
	X_1	1	1	1	1	1	3	2	4
WEV	E_2	0	0	1	1	-2	-2	1	1
	X_2	1	1	2	2	0	0	4	4

- WEV errors do *not* balance on any individual
- WEV errors do *not* balance on any fixed value of T

Incorrect characterizations of the CTT in the literature

In *Applied Measurement in Education*, 2010:

> *One can define T as the expected value of the observed scores X, which leads to the expected value of E being zero. [Or], one can define the expected value of E as zero, which leads to T being the expected value of X.*

- This specifies the WCE model
- WCE is the *weakest* of the 6 MEMs
- Lacks balance of errors on *every individual*
- Does not ensure true scores and errors are *uncorrelated*.

CTT vs. WCE

	Test/Retest	1	2	1	2	1	2	1	2
	Individual	1	1	2	2	3	3	4	4
	T	1	1	1	1	2	2	3	3
CTT	E_1	0	0	0	0	-1	1	-1	1
	X_1	1	1	1	1	1	3	2	4
WCE	E_2	1	1	1	0	0	-1	-1	-1
	X_2	2	2	2	1	2	1	2	2

- WCE errors do *not* balance on any individual or true scores
- WCE errors can correlate with T

In summary

- Several equivalent ways to define the CTT model (see paper).
- The exchangeability captured by the

$$T = \mathbb{E}(X \mid \sigma(i))$$

  condition is very strong: *errors balance on every individual*.
- This condition is the *defining feature* of the CTT.
- Without it, we have *no ability* to make reasonable inferences down to the level of the test-taking individual.

Thank you!

Edward Kroc
Assistant Professor of Measurement, Evaluation, and Research Methodology
Department of ECPS, UBC

Office: Scarfe 2526
Phone: (604)-822-8671
Email: ed.kroc@ubc.ca
IG: @ed_kroc
Website: ekroc.weebly.com