Center for Advanced Studies in Measurement and Assessment

CASMA Technical Note

Number 6

Standard Errors of Measurement of the Difference in Group Mean Difficulty Levels for Items

Robert L. Brennan[1]

March 17, 2013

[1]Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, Center for Advanced Studies in Measurement and Assessment (CASMA).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

Abstract

This paper considers standard errors of measurement of the difference in group mean difficulty levels for items. It is assumed that the items are the same for the two groups, but the persons are different. The basic question considered here is, "What is the variability of such group mean differences over replications?" This question is ambiguous, however, without specifying what constitutes a replication. There are at least two possibilities. First, each replication could consist of a *different* sample of items administered to the *same* persons. Second, each replication could consist of a *different* sample of items administered to a *different* sample of persons. These two cases are treated, followed by a consideration of the conditional standard error of measurement for differences in the difficulty level of a single item for two groups. Multivariate generalizability theory provides the theoretical framework and statistical machinery for most of the issues addressed here.

This paper treats standard errors of measurement ($SEMs$) of the difference in group mean difficulty levels for items. For purposes of specificity, consider the highly synthetic data in Table 1 at the bottom of this page. This table contains responses for 10 dichotomously-scored items for 6 examinees in group a and 12 examinees in group b. It is important to note that the items are the same for the two groups, but the persons are different.

For the fourth item, for example, the difficulty level for group a is .1667 and the difficulty level for group b is .4167. The mean difficulty level for group a is .3333, the mean difficulty level for group b is .5833 and, therefore, the difference in group mean difficulty levels is $.3333 - .5833 = -.2500$. The basic question considered in this paper is, "What is the variability, over replications, of the difference in group mean difficulty levels for items?"[2]

This question is ambiguous, however, without specifying what constitutes a replication. There are at least two possibilities. First, each replication could consist of a *different* sample of items administered to the *same* persons, which we will call the "items random" case. Second, each replication could consist of a *different* sample of items administered to a *different* sample of persons, which we will call the "items and persons random" case. These two cases are treated next, followed by a consideration of the conditional standard error of

[2] Note that the difference in group mean difficulty levels for items is identical to the difference in group mean proportion-correct scores for persons. Therefore, the $SEMs$ will be the same *provided* the data collection design is the same—items are the same for the two groups, and persons are different for the two groups.

Table 1: Synthetic Data

	Group a							Group b												
Item	Scores (X_{ipa})						$\overline{X}_{ia}$	Scores (X_{ipb})												$\overline{X}_{ib}$
1	1	0	0	0	0	0	.1667	1	0	1	0	0	0	0	0	0	0	0	0	.1667
2	0	0	0	0	0	0	.0000	1	1	1	0	0	1	0	0	0	0	0	0	.3333
3	0	0	0	0	0	0	.0000	1	1	1	1	1	0	0	0	0	0	0	0	.4167
4	0	1	0	0	0	0	.1667	1	1	0	1	1	0	0	1	0	0	0	0	.4167
5	1	0	0	0	0	0	.1667	1	1	1	1	1	0	1	0	0	0	0	0	.5000
6	1	0	0	0	0	0	.1667	1	1	1	0	1	1	1	0	0	0	0	0	.5000
7	1	0	0	0	0	0	.1667	1	1	1	1	1	1	1	0	0	0	0	0	.5833
8	1	1	1	1	0	0	.6667	1	1	1	1	0	1	1	1	1	1	0	0	.7500
9	1	1	1	1	1	0	.8333	1	1	1	1	1	1	1	1	1	1	1	0	.9167
10	1	1	1	1	1	1	1.0000	1	1	1	1	1	1	1	1	1	1	1	1	1.0000
Mean							.3333													.5583
S.D.							.3600													.2607

Note. The covariance of item difficulty levels is $\widehat{S}_i(\overline{X}_{ia}, \overline{X}_{ib}) = .0849$ and the correlation of item difficulty levels is $.0849/[(.3600)(.2607)] = .904$.

measurement for differences in the difficulty level of a single item for two groups. Multivariate generalizability theory provides the theoretical framework and statistical machinery for most of the issues addressed here, particularly those in Sections 2 and 3.

1 Items Random

This section uses the following notational conventions:

- X_{ipa} = observed score for item i and person p in group a;
- $\overline{X}_{ia}$ = observed mean for item i in group a (i.e., difficulty level for item i in group a);
- $\overline{X}_{a}$ = mean over items and persons in group a (i.e., mean of item difficulty levels in group a);

Corresponding conventions apply to X_{ipb}, $\overline{X}_{ib}$, and $\overline{X}_{b}$

If items are random, but not persons (i.e., persons are fixed), then by standard statistical results

$$SEM((\overline{X}_a - \overline{X}_b)|P) = \sqrt{S^2(\overline{X}_a) + S^2(\overline{X}_b) - 2\,S(\overline{X}_a, \overline{X}_b)}, \tag{1}$$

where "$|P$" designates that persons are fixed, $S^2(\star)$ stands for variance and $S(\star,\star)$ stands for covariance. It is well known that the variance of a mean equals the variance of the individual elements (items, in this case) divided by the number of them (n_i, in this case). A similar result holds for covariances; i.e., in this case, $S(\overline{X}_a, \overline{X}_b)) = S(\overline{X}_{ia}, \overline{X}_{ib})/n_i$. It follows that

$$SEM((\overline{X}_a - \overline{X}_b)|P) = \sqrt{\frac{S_i^2(\overline{X}_{ia})}{n_i} + \frac{S_i^2(\overline{X}_{ib})}{n_i} - 2\,\frac{S_i(\overline{X}_{ia}, \overline{X}_{ib})}{n_i}}, \tag{2}$$

where the subscript i in $S_i^2(\star)$ and $S_i(\star,\star)$ indicates that the variances and covariance are taken over items.

An unbiased estimator of $SEM((\overline{X}_a - \overline{X}_a)|P)$ is obtained by using unbiased estimators of each of the variances and the covariance in Equation 2. For the data in Table 1,

$$\widehat{SEM}((\overline{X}_a - \overline{X}_b)|P) = \sqrt{\frac{.3600^2}{10} + \frac{.2607^2}{10} - 2\left(\frac{.0849}{10}\right)} = .0528. \tag{3}$$

2 Items and Persons Random

Multivariate generalizability (G) theory (Brennan, 2001, chaps 9 and especially 10) is ideally suited to consider the case of both items and persons random. In the notational conventions of multivariate G theory, the design in Table 1 is

$i^{\bullet} \times p^{\circ}$; i.e., there are two levels of a fixed facet (group a and group b), all items are taken by persons in both groups ($i^{\bullet}$), and the persons in the two groups are different (p°). The SEM for both items and persons random is a D study statistic associated with this design:

$$SEM(\overline{X}_a - \overline{X}_b) = \sqrt{\sigma^2(\overline{X}_a) + \sigma^2(\overline{X}_b) - 2\,\sigma(\overline{X}_a, \overline{X}_b)}, \tag{4}$$

where

$$\sigma^2(\overline{X}_a) = \frac{\sigma_a^2(i)}{n_i} + \frac{\sigma_a^2(p)}{n_a} + \frac{\sigma_a^2(ip)}{n_i\, n_a}, \tag{5}$$

$$\sigma^2(\overline{X}_b) = \frac{\sigma_b^2(i)}{n_i} + \frac{\sigma_b^2(p)}{n_b} + \frac{\sigma_b^2(ip)}{n_i\, n_b}, \tag{6}$$

and

$$\sigma(\overline{X}_a, \overline{X}_b) = \frac{\sigma_i(\overline{X}_{ia}, \overline{X}_{ib})}{n_i} = \frac{S_i(\overline{X}_{ia}, \overline{X}_{ib})}{n_i}. \tag{7}$$

Note that n_a is the number of persons for group a, and n_b is the number of persons for group b.[3]

The numerators of the terms to the right of the equal sign in Equation 5 [namely, $\sigma_a^2(i)$, $\sigma_a^2(p)$, and $\sigma_a^2(ip)$] are G study variance components for a univariate $i \times p$ design for group a. They are easily estimated using mean squares. Specifically,

$$\begin{aligned}
\widehat{\sigma}_a^2(i) &= \frac{MS_a(i) - MS_a(ip)}{n_a}, \\
\widehat{\sigma}_a^2(p) &= \frac{MS_a(p) - MS_a(ip)}{n_i}, \quad \text{and} \\
\widehat{\sigma}_a^2(ip) &= MS_a(ip).
\end{aligned}$$

Replacing a with b in the above equations gives $\widehat{\sigma}_b^2(i)$, $\widehat{\sigma}_b^2(p)$, and $\widehat{\sigma}_b^2(ip)$ in Equation 6. Note that the covariance component denoted $\sigma_i(\overline{X}_{ia}, \overline{X}_{ib})$ in Equation 7 is identical to the covariance denoted $S_i(\overline{X}_{ia}, \overline{X}_{ib})$ in Equation 2.[4]

It can be shown that

$$SEM(\overline{X}_a - \overline{X}_b) = \sqrt{SEM^2((\overline{X}_a - \overline{X}_b)|P) + \frac{\sigma_a^2(p)}{n_a} + \frac{\sigma_b^2(p)}{n_b}}. \tag{8}$$

This means that $SEM(\overline{X}_a - \overline{X}_b) \geq SEM((\overline{X}_a - \overline{X}_b|P))$, as must be the case since there are two sources of variability (items and persons) for $SEM(\overline{X}_a - \overline{X}_b)$ and only one (items) for $SEM((\overline{X}_a - \overline{X}_b)|P)$. Obviously, the two $SEMs$ get closer in magnitude as n_a and n_b get larger.

[3] There is a slight notational inconsistency in Equations 5 and 6, namely, the group identifier is a subscript of $\overline{X}$ to the left of the equal signs, whereas the group identifier is a subscript of σ^2 to the right of the equal signs. The latter is consistent with the conventions for multivariate G theory in Brennan (2001); the former is more consistent with introductory statistical texts.

[4] This equality of a covariance component and an observable covariance occurs when there is only one linked facet (identified with a $\bullet$) in a multivariate design.

For the synthetic data in Table 1, it can be shown that

$$\widehat{SEM}(\overline{X}_a - \overline{X}_b) = \sqrt{\begin{array}{c} \left[\frac{.11407}{10} + \frac{.03333}{6} + \frac{.09333}{(10)(6)}\right] \\ + \left[\frac{.05741}{10} + \frac{.07542}{12} + \frac{.12685}{(10)(12)}\right] \\ -2\left(\frac{.08488}{10}\right) \end{array}} = .1209 \qquad (9)$$

The Appendix provides mGENOVA (Brennan, 2001b) control cards to obtain $\widehat{SEM}(\overline{X}_a - \overline{X}_b)$ for the data in Table 1.[5] The SEM is part of the D study output identified as

"`Composite Error Standard Deviation for Mean`".

The estimates of the variance components and the covariance component in Equations 5–7 are part of the G study output labeled

"`ESTIMATED G STUDY VARIANCE AND COVARIANCE COMPONENTS`".

The estimated variance components can be obtained also using the mean squares in the G study output labeled

"`STATISTICS FOR ESTIMATING G STUDY VARIANCE AND COVARIANCE COMPONENTS`".

Alternatively, two separate runs of GENOVA (Crick & Brennan, 1983) can be used to estimate $\sigma^2(\overline{X}_a)$ and $\sigma^2(\overline{X}_b)$, respectively, and the covariance of the item difficulty levels can be computed separately.

3 Persons Random and *CSEMs*

The SEM for the difference in difficulty levels for a *single* item is

$$CSEM_i = \sqrt{\frac{S_p^2(X_{pia})}{n_a} + \frac{S_p^2(X_{pib})}{n_b}}, \qquad (10)$$

where $CSEM$ stands for "conditional" SEM. There is no covariance term in Equation 10 because $CSEM_i$ is for one item only. Note that $S_p^2(X_{pia}/n_a$ is the variance of the mean (i.e., the difficulty level) for samples of persons from group a, and $S_p^2(X_{pib})/n_b$ is the variance of the mean (i.e., the difficulty level) for samples of persons from group b. Since $CSEM_i$ involves sampling persons, we say that persons are random. Clearly, items are fixed at a sample size of one, since we focus on each item separately and do not generalize to some larger set of items.

[5]Currently, mGENOVA is dimensioned such that the total number of persons in the two groups cannot exceed 999.

An estimator of $CSEM_i$ is

$$\widehat{CSEM}_i = \sqrt{\frac{\widehat{S}_p^2(X_{pia})}{n_a} + \frac{\widehat{S}_p^2(X_{pib})}{n_b}}, \tag{11}$$

where $\widehat{S}_p^2(\star)$ denotes an unbiased estimate of the variance (over persons) of the item scores for a particular group. When items are scored dichotomously,

$$\widehat{CSEM}_i = \sqrt{\frac{\overline{X}_{ia}(1-\overline{X}_{ia})}{n_a - 1} + \frac{\overline{X}_{ib}(1-\overline{X}_{ib})}{n_b - 1}}. \tag{12}$$

As discussed by Brennan (2001a, sect. 5.4), the two terms in the square root of Equation 12 are Lord's (1955, 1957) conditional absolute error variances.[6]

Consider, for example, the fourth item in Table 1. For group a the difficulty level is .1667, for group b the difficulty level is .4167, and using Equation 12,

$$\widehat{CSEM}_i = \sqrt{\frac{.1667(1-.1667)}{6-1} + \frac{.4167(1-.4167)}{12-1}} = .2233. \tag{13}$$

If the mGENOVA `DOPTIONS` control card contains `CSEM` (see the Appendix), a distribution of values of $CSEM_i$ is provided.

4 Comments

The basic question considered in this paper is, "What is the variability of differences in group mean difficulty levels ($\overline{X}_a - \overline{X}_b$) over replications?" This question has different answers depending on how replications are defined. The SEM is necessarily smaller when only items are random than when both items and persons are random. In the author's opinion, assuming both items and persons are random usually seems more reasonable, although the computations are more complicated. (The computational complexity disappears using mGENOVA.)

Consider the sample sizes in the denominators in Equation 2 for items random, Equations 5–7 for both persons and items random, and Equation 11 for $CSEM_i$ with persons random. Strictly speaking, these denominators are D study sample sizes that need not equal the same sizes in the G study data (e.g., Table 1). The sample sizes in these denominators can take on any value(s) desired by the investigator. When the G and D study sample sizes differ, the estimated $SEMs$ (or $CSEM_i$) provide projections of what these statistics would be for the D study sample sizes. (Essentially, this is the G theory extension of the Spearman-Brown formula in classical test theory.) [7]

The discussion in this paper has been in terms of traditional difficulty levels based on dichotomous data such as that in Table 1. The equations, however, have no such restriction, except for Equation 12. So, $\overline{X}_{ia}$ and $\overline{X}_{ib}$ are, more generally item mean scores over persons in groups a and b, respectively.

[6] Lord presented his results for person scores, rather than the item scores (difficulty levels) considered here.

[7] Note that Equation 12 assumes that the G and D study sample sizes are the same.

5 References

Brennan, R. L. (2001a) *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L. (2001b). *mGENOVA*. Iowa City, IA: Iowa Testing Programs, University of Iowa. [Computer software and manual.] Available on www.education.uiowa.edu/casma

Crick, J. E. & Brennan, R. L. (1983). *GENOVA: A generalized analysis of variance system* (American College Testing Technical Bulletin No. 43). Iowa City, IA: ACT, Inc. [Computer software and manual.] Available on www.education.uiowa.edu/casma

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, *15*, 325–336.

Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, *17*, 510–521.

6 Appendix: mGENOVA Control Cards

```
GSTUDY   SEM for difference in group mean difficulty levels
OPTIONS  NREC 10  "*.out"
MULT     2 Group-a  Group-b
EFFECT   * i    10 10
EFFECT     p     6 12
FORMAT   7 0
PROCESS
Item  1    1  0  0  0  0  0    1  0  1  0  0  0  0  0  0  0  0  0
Item  2    0  0  0  0  0  0    1  1  1  0  0  1  0  0  0  0  0  0
Item  3    0  0  0  0  0  0    1  1  1  1  1  0  0  0  0  0  0  0
Item  4    0  1  0  0  0  0    1  1  0  1  1  0  0  1  0  0  0  0
Item  5    1  0  0  0  0  0    1  1  1  1  1  0  1  0  0  0  0  0
Item  6    1  0  0  0  0  0    1  1  1  0  1  1  1  0  0  0  0  0
Item  7    1  0  0  0  0  0    1  1  1  1  1  1  1  0  0  0  0  0
Item  8    1  1  1  1  0  0    1  1  1  1  0  1  1  1  1  1  0  0
Item  9    1  1  1  1  1  0    1  1  1  1  1  1  1  1  1  1  1  0
Item 10    1  1  1  1  1  1    1  1  1  1  1  1  1  1  1  1  1  1
DSTUDY   SEM for difference in group mean difficulty levels
DOPTIONS NEGATIVE CSEM
WWTS     1  -1
DEFFECT  $ i 10 10
DEFFECT    P  6 12
ENDDSTUDY
```