Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 26 Number 14, June 2021 ISSN 1531-7714

Conditional Standard Error of Measurement: Classical Test Theory, Generalizability Theory and Many-Facet Rasch Measurement with Applications to Writing Assessment

Alan Huebner, *University of Notre Dame*
Gustaf B. Skar, *Norwegian University of Science and Technology*

Writing assessments often consist of students responding to multiple prompts, which are judged by more than one rater. To establish the reliability of these assessments, there exist different methods to disentangle variation due to prompts and raters, including classical test theory, Many Facet Rasch Measurement (MFRM), and Generalizability Theory (G-Theory). Each of these methods defines a standard error of measurement (SEM), which is a quantity that summarizes the overall variability of student scores. However, less attention has been given to conditional SEMs (CSEM), which expresses the variability for scores of individual students. This tutorial summarizes how to obtain CSEMs for each of the three methods, illustrates the concepts on real writing assessment data, and provides computational resources for CSEMs including an example of a specification file for the FACETS program for MFRM and R code to compute CSEMs for G-theory.

Introduction

Writing assessments are used by various national agencies in many countries to monitor the development of students' writing proficiency and development. In writing assessment, major threats to reliability are rater and task effects and their interaction. A rater is anyone with the responsibility to judge the quality of a student text on the basis of some criteria. A writing task can take an infinite number of shapes, but it is common that a task prompts a student to write in a particular genre, which can be specified in terms of the function the writing serves (e.g., informative, argumentative, descriptive) or by labels such as "letter to the editor", "cooking recipe", "expository essay", or by a combination (e.g., "write an argumentative text as a letter to the editor"). Writing assessment research has consistently found rater effects to be non-trivial (Eckes, 2015), as raters within and across contexts (e.g., school districts, countries) differ in their judgement of text quality. The hitherto only international writing assessment investigation ended in an anticlimax, as raters in different countries were found to disagree on the merits of texts (Purves, 1992). The task effect in writing assessment has also proven to be substantial. A study by Bouwer et al. (2015) took into account both raters and tasks and suggested that students need to write at least 12 texts (three in each of four genres), rated by at least two independent raters to increase reliability to an acceptable level.

In many contexts, multiple tasks and multiple raters are costly, and it is unfeasible to include 12 tasks. When basing decisions on the outcomes of writing assessment, a decision maker can be aided by estimates of the uncertainty of measures. The *standard error of measurement* (SEM; Harville, 1991) can be used to estimate a confidence interval at a given level (e.g., 68 %, 95 %, 99%) around the observed student score by multiplying the SEM by 1, 1.96 or 2.58, and then subtracting and adding that value to the observed score. This confidence interval is useful in situations in which cut scores are

used, as it can help the practitioner assess the impact of establishing cut scores at certain levels. It can also be helpful to researchers wishing to include texts representing distinct proficiency levels. This SEM may be referred to as a "general" SEM, as it is a fixed value used for all candidates. This means that the confidence interval is equal in size across students and across different levels of competence. On the other hand, the *conditional SEM* (CSEM) differs by taking into account that the standard error of measurement is not a score-invariant property (Embretson & Reise, 2000; Feldt et al., 1985).

Reliability and SEM in Writing Assessment

In writing assessment, reliability is often used interchangeably with inter-rater agreement, which denotes the extent to which two or more raters agree on the judgement of a piece of writing. Intra-rater agreement, which to our knowledge is less commonly investigated, refers to the extent to which a single rater is consistent over time. In a comprehensive review of methods for establishing inter-rater agreement, Stemler (2004) distinguished between consensus, consistency, and measurement approaches. Consensus approaches involve calculations that derive an index of proportions of exact or adjacent agreement, while consistency approaches are concerned with the consistency in rank ordering students. Consensus and consistency approaches were developed under the *classical test theory* (CTT) paradigm. Measurement approaches can be used to both estimate effects and disentangle additional effects (e.g., task effects, effects of writing at different occasions). Some measurement approaches belong to the modern test theory paradigm. Two of the methods mentioned by Stemler (2004) as measurement approaches are *Generalizability Theory* (G-theory; Brennan, 2001) and *Many-Facet Rasch Measurement* (MFRM; Linacre, 1994). Both methods represent more sophisticated techniques and can be used to disentangle multiple sources of variation in a writing assessment context, albeit under very different premises.[1]

The general SEM and the G-theory and MFRM approaches have received outstanding treatments in several papers and books. Harville (1991) offers a good starting point for understanding SEM under CTT approaches, G-theory has been described by Shavelson and Webb (1991) and Brennan (2001), and MFRM and other versions of the Rasch model have been thoroughly described in, for example, Bond and Fox (2015), Eckes (2015), Linacre (1994), and McNamara (1996). These approaches have also received due attention in the language testing field where there have been several comparisons between CTT, MFRM, and G-theory (e.g., Bachman, 2004). Lynch and McNamara (1998) compared G-theory and MFRM in terms of analysis for test development purposes, and Sudweeks et al. (2005) made comparisons of the two methods on several criteria, including interpretation of results and handling of missing data. Recently, a comprehensive introduction to these and other methods were presented in Aryadoust and Raquel (2019), accompanied by tutorials to conduct several relevant analyses.

In all contexts in which students' writing is measured for decision making, it is common to report the reliability of the measures and leave to the user of the results to appraise the trustworthiness of students' results. Despite a rich literature, we have found very few resources dealing with SEM and CSEM for writing assessment, although SEMs offer a practical tool for assessing the appropriateness of, for example, dividing students into groups of pass and fail. For G-theory, there are accounts by Brennan (e.g., 1998, 2001), but these do not include tutorials on the procedures using widespread statistical software, such as SPSS and the R statistical computing environment (R Core Team, 2019). Moreover, there are very few illustrations of G-theory CSEM on real data, outside of the work of Brennan. For MFRM, various instructional papers and chapters tend to focus more on the so-called fit statistic. This helps the researcher to gauge to what extent a person's responses fit the MFRM model but is less helpful for establishing confidence intervals around person scores. To add to the literature, this article describes some approaches to SEM using, in Stemler's terms, consistency measures for estimating a general SEM and measurement methods for estimating CSEM. The intention of this broad approach is to offer some initial guidance to researchers working either mainly with classical test theory approaches or with measurement approaches.

[1] In fact, the CTT, G-theory and Rasch approaches are often said to represent different philosophies. It is beyond the scope of this practical guide to review these differences, but interested readers are referred to Embretson and Reise (2000) and Brennan (2001) for detailed accounts.

This Tutorial

In this tutorial, we present analyses conducted on real writing assessment data with the CTT[2], MFRM, and G-theory approaches, respectively. We describe how to estimate the general SEM for the CTT approach and CSEM in MFRM and G-theory, with the aim of enabling other researchers to gain familiarity with the procedures. Specifications for the software used are provided in two appendices. The remainder of this paper is structured as follows. The Methods section briefly recaps CTT, MFRM, and G-theory as well as the computational resources available for obtaining CSEMs in each of these methods. The Real Data Analysis section describes data from a study evaluating rater training for a writing assessment program in Norway and illustrates the basic results and CSEMs from the three methods above. The Conclusion includes comments on the methods as well as ideas for further investigation.

Methods

CTT Concepts and Computational Resources

Under the traditional CTT approach a reliability estimate is calculated, and the estimate is then, alongside the standard deviation for the test score, plugged into the following equation:

$$SEM = S_X\sqrt{1 - r_{xx}} \ ,$$

where SEM is the standard error of measurement, S_X is the standard deviation and r_{xx} is the reliability estimate. Traditionally, r_{xx} has represented the correlation between parallel test forms. For writing assessment, however, a test is not a single entity. The difficulty of holding task difficulty and rater severity constant are two reasons for this; a third is that in practice, especially in large scale assessments, all raters will not judge all student texts. Students will therefore encounter different tests depending on whom judges the text (see chapter 2 of McNamara [1996], for a treatment of these aspects), and depending on which prompt they chose in settings when it is possible to choose among tasks.

A common way to estimate reliability when raters perform judgements of the qualities of students' responses is to compute a consistency measure, for example Spearman's rho or Kendall's tau (Stemler, 2004). Another popular estimate is the intraclass correlation coefficient (ICC; McGraw & Wong, 1996). These consistency measures can be interpreted as an expression of the degree to which to raters coincide in their judgement. A strong correlation indicates that raters consistently award the same candidates with high scores and the same candidates with low scores. McNamara (2000) proposes that a correlation of .7 is a "rock-bottom minimum" for language tests. It should be noted that the consistency measure does not indicate to which extent raters agree on the exact score, and so the measure will indicate if raters are ranking student responses in similar or dissimilar fashion rather than if they award the same mark. There are several other popular statistics that can either produce estimates of absolute agreement or consistency, thereby complementing the correlation (e.g., Stemler, 2004; Kilem, 2014), some of which are incorporated in R packages such as *CTT* (Willse, 2018) and *psych* (Revelle, 2019).

MFRM Concepts and Computational Resources

MFRM (Linacre, 1994) is an extension of the Rasch model (Rasch, 1980). The latter is a probabilistic model stating that a student's probability of scoring correct (or affirmative) on a dichotomous item is equal to the difference between the student's modelled ability and the particular item's modelled difficulty. Formally, the Rasch model usually takes this expression:

$$log\left(\frac{P_{ni}}{1 - P_{ni}}\right) = B_n - D_i \ ,$$

where P_{ni} is the probability of a correct response, B_n denotes the ability of student n, and D_i denotes the difficulty of item i. When the ability is equal to the difficulty, a student has a probability of .5 of responding correctly. Estimates of student ability and item difficulty are expressed on a logit scale, and it is common for student ability to be in the range of -5 to 5.

In contrast to CTT approaches and, as we shall see, the G-theory approach, applying the Rasch model is a means to scale student scores. If the empirical data fits

[2] There are methods for estimating CSEM under the CTT approach (Feldt et al., 1985), but we have not yet encountered descriptions of how to do so with data stemming from judgements of student texts (or any other artifacts).

the assumptions of the Rasch model, the scaling transforms the ordinal scores to interval scores (Engelhard, 2013). Another feature of fitting scores to the Rasch model is that the resulting estimates are invariant of each other. A student thus has an ability level invariant of rater, item and task, and an item has a difficulty level invariant of the other facets. There are several outstanding treatments of the Rasch model in educational and language testing (e.g., Bond & Fox, 2015; McNamara, 1996), and readers are advised to refer to them for additional technical specifications.

The MFRM is an extension of the Rasch rating scale model (which is another extension of the Rasch model; Andrich, 2016), used in educational and psychological testing where persons perform tasks that are judged into one of k categories.[3] It does so by treating aspects of the measurement causing variation as "facets," which in turn comprise elements. In the MFRM terminology, student groups, raters, tasks, and rating scales are all facets, and individual persons or items in these groups are referred to as elements.

A MFRM model can take many shapes, depending on the purpose of the analysis. In the present case we have used the following model:

$$log\left(\frac{P_{nmijk}}{1-P_{nmijk}}\right) = B_n - D_m - E_i - C_j - F_x\,,$$

where P_{nmijk} represents the probability of student n on task m, rating scale i, by rater j receiving a score of k, and $P_{nmijk} - 1$ represents the probability of the same student under the same conditions receiving a score of $k-1$. B_n is the ability for person n, D_m is the difficulty of task m, E_i is the difficulty of rating scale i, and C_j is the severity of rater j. Finally, F_x represents the point on the logit scale where category k and $k-1$ are equally probable.

Focusing particularly on the precision of measures, there are a few key statistics ensuing from a MFRM analysis. A prerequisite, however, for using the output is that the data fits the model. The MFRM software we have used (FACETS) does not output a meaningful global measure of data-model fit, but, as Eckes (2015, p. 69) notes, the global fit can easily be assessed using standardized residuals. Eckes suggest that when there are less than 5 % standardized residuals exceeding $|2.0|$ and 1 % exceeding $|3.0|$ the data fits the model reasonably well. It can also be noted that MFRM software outputs element-specific fit measures, indicating to what extent the model has been able to predict an element's raw scores. These are called information-weighted fit (or *infit*) and outlier-sensitive fit (or *outfit*). Fit statistics exceeding 1.0 indicate "misfit," or unpredictable differences between observed and modelled expected results, while fit statistics below 1.0 indicate "overfitting" elements (i.e. elements with less than expected variation). For a non-technical treatment of these statistics, see Linacre (2002, p. 878). When the researcher has concluded that the data fits the model reasonably well, there are four precision measures to take into consideration. The first three are versions of a global reliability estimate. Building on Schumacker and Smith (2007, p. 399), we will briefly present them here. The separation statistic R, which for the persons facet is a Rasch analogue to coefficient alpha , indicates to what extent elements of facets have been reliably separated. The R value can be interpreted as the ability of the measurement to reliably separate candidates. Using the person facet as an example, R is calculated by dividing the person facet variance (SD_p^2) from SA_p^2, which is a "person variance that is adjusted for measurement" (Schumacker & Smith, 2007, p. 399). The latter is calculated by subtracting the mean square error (MSE_p) from the variance. MSE_p is given by

$$MSE_p = \sum_{n=1}^{N} S_n^2/N\,,$$

where S_n^2 is the standard error for each person. R has a maximum value of 1.0 and can be converted to "separation statistic" (G_p), by $SA_p/\left(MSE_p\right)^{1/2}$ which can be interpreted as the ability for a particular configuration of persons, raters and task to separate persons. Unlike R, G does not have a maximum value. G_p can, in turn, be converted to H_p, which indicates the number of significant "strata" a particular measurement

[3] It is also possible to specify a Partial Credit Model (PCM) using the FACETS software. PCM is particularly useful when items do not share number of scale steps, or when one wishes to investigate if items behave differently (see Eckes, 2014, pp. 127–132.

can divide students into. H_p is given by $(4G_p + 1)/3$. It follows that the MFRM analysis outputs conditional SEMs, i.e. a SEM for each element. The SEM for subject n is calculated this way:

$$S_n^2 = \frac{1}{\sqrt{TI(B_n)}},$$

where $TI(B_n)$ is the test information for person n, with ability level B. The test information is computed by summing the model variance for each element. SEMs, reliability, and separation measures are also given for other facets than the student facet.

With reference to situations where a researcher is forced to make absolute decisions, the separation statistics are useful for estimates of the measures ability for the relative separation of the persons (or other elements), while the SEM provides the researcher with means to gauge possible classification errors, after a cut score has been established.

Conducting a MFRM Analysis

There are several programs that enable the researcher to perform MFRM analyses. Freeware developed for R include the *SIRT* (Robitzsch, 2020) and *TAM* (Robitzsch et al., 2020) packages, and RUMM® is a commercial package. A popular program is FACETS®, developed and maintained by Linacre (e.g., 2018) for the past 30 years. In this tutorial, we provide examples for specifying settings for the FACETS software (see Appendix A) and how to read and make use of some of the output, with a particular focus on SEM.

To conduct a MFRM analysis in FACETS, one needs to create a "specification file." As FACETS is a versatile tool, there is an abundance of methods to specify the analysis, all depending on how to best accommodate the researcher's need. In this example, we have specified a relatively simple analysis. FACETS operates with "centered" and "non-centered" (or "floating") facets. In the context of the program, a centered facet is the "local origin," while a non-centered facet is, as it were, floating in relationship to that origin. This means that all facets, except one, are centered so that the average logit value for a facet is 0. The non-centered facet is measured against this origin and may thus have a positive or negative mean. For non-technical audiences a mean score of 0 introduces interpretational difficulties, since some candidates will have "negative ability" and some raters "negative severity". To accommodate reporting needs, the FACETS specification file can be amended with a user specified mean score. For this data set, we have set the mean score to 50, which is conventional in some educational assessment contexts, and the distance between each logit to 10 (see Appendix A for instruction on how to enter this specification).

G-theory Concepts and Computational Resources

G-theory is a framework for quantifying reliability in which sources of variation are also referred to as "facets." Unlike in the MFRM context, the objects of measurement (often people) are not considered a facet in G-theory; on the other hand, factors such as items, raters, and occasions are regarded as facets. G-theory analyses are described as having two phases, the G study and the D study. In the G study phase, estimates are obtained for variance components of the facets and interactions between them, so that the largest sources of variation can be identified. Then, the D study provides coefficients of overall reliability and also allows the researcher to obtain projections of the reliability levels yielded by sample sizes different than the ones used for the study. Thus, the D study allows practitioners to determine procedures for optimal research designs. Furthermore, G and D studies may be conducted for many different experimental designs, as facets may be crossed or nested within each other. Also, G-theory accommodates facets as being either random (e.g., the researcher wishes to generalize beyond the particular sample of raters used in the study) or fixed (e.g., the researcher does not wish to generalize beyond those particular raters).

While many previous papers have applied G and D study methodologies to a number of different fields in social and biomedical science, there is a relative scarcity of studies demonstrating the calculations of G-theory CSEMs. A notable exception is Brennan (1998), who derives CSEMs for a number of different G-theory designs and presents examples for a dichotomously-scored vocabulary test and a polytomously-scored mathematics assessment. Brennan (2001a) also presents these examples as well as a summary of the concepts and calculations for G-theory CSEMs.

We briefly review G-theory concepts and notation to facilitate the discussion of CSEMs. See Shavelson and

Webb (1991) and Brennan (2001a) for book-length treatments on fundamental and advanced concepts in G-theory. The persons which are the subject of measurement are denoted as p, and facets such as items, raters, and occasions are denoted as i, r, and o, respectively. The variance component for persons is denoted as $\sigma^2(p)$, the variance component for the persons by items interaction is denoted as $\sigma^2(pi)$, and so on. The actual sample size for items is notated as n_i, and so on for other facets. Then, n_i' denotes the sample size for items considered in the D study, which is not necessarily equal to n_i, and so on for the other facets.

In a random model, the particular raters, occasions, etc. are considered to be drawn from a very large (infinite) groups of raters, occasions, etc. A universe score for person p, notated as μ_p, is her or his average score over all items, raters, and occasions we wish to generalize to. The absolute error variance $\sigma^2(\Delta)$ for designs with random facets is the sum of all the variance components except the subject variance $\sigma^2(p)$. For example, consider a design with two crossed random facets, items and occasions, which is denoted as $p \times i \times o$. Then, the absolute error variance for this design is given by

$$\sigma^2(\Delta) = \frac{\sigma^2(i)}{n_i'} + \frac{\sigma^2(o)}{n_o'} + \frac{\sigma^2(pi)}{n_i'} + \frac{\sigma^2(po)}{n_o'} + \frac{\sigma^2(io)}{n_i' n_o'} + \frac{\sigma^2(pio)}{n_i' n_o'}.$$

The absolute error for person p is defined as

$$\Delta_p \equiv \bar{X}_p - \mu_p ,$$

which is interpreted as the error resulting from using the observed mean score from person p as an estimate of that person's universe score (Brennan, 2001a). Then, in the design above, the absolute CSEM is defined as

$$\hat{\sigma}(\Delta_p) = \sqrt{\frac{\hat{\sigma}^2(i)_p}{n_i'} + \frac{\hat{\sigma}^2(o)_p}{n_o'} + \frac{\hat{\sigma}^2(io)_p}{n_i' n_o'}},$$

Thus, $\sigma(\Delta_p)$ can be described as the standard error of the *within*-person mean, and the average of squares of the $\hat{\sigma}(\Delta_p)$ values is equal to the absolute error variance, $\sigma^2(\Delta)$ (Brennan, 2001a).

To our knowledge, the resources for obtaining absolute CSEMs for G-theory are not numerous. The freely available software mGENOVA (Brennan, 2001b) includes an option that, when called, will supply the CSEMs for a number of different G-theory designs. However, this is a standalone software, and in a modern data science context R offers the advantage of being able to handle all steps of the analysis, including data cleaning, analysis, visualization, and reporting. Thus, we provide R code in Appendix B for computing absolute CSEMs for G-theory, building upon the G and D study capabilities of the *gtheory* package (Moore, 2016).

Real Data Analysis

Data

The data used in the current paper was collected in the fall of 2016 from an investigation conducted to evaluate a rater training approach. The context of rater training was the Norwegian Sample-Based Writing Test (NSBWT) which was a government-financed writing test program measuring writing proficiency among students in 5th and 8th grades (Skar, 2017; Jeffery et al., 2018). Raters (N = 8) from the NSBWT were sampled and rated texts from 25 students enrolled in 8th Grade in Norway. Each student had written two texts, representing two genres, and each text was assessed on six rating scales, *Communication*, *Content*, *Text Structure*, *Language Use*, *Spelling*, and *Punctuation*, each scored from 1 to 5. The resulting dataset comprised 2,400 ratings and the design was completely crossed, as all raters judged both genres from all students. The results, which were based on MFRM, were published in Skar and Jølle (2017) and used as part of the quantitative evaluation. For more information about the rating scales and the training and evaluation, readers are referred to Skar and Jølle (2017).

CTT Results

We estimated reliability using Spearman's rho, which takes into account that the data is expressed on an ordinal scale, rather than an interval scale (Stevens, 1946). For the present data set, the average correlation was rho = .62; Table 1 presents all inter-rater correlations. Coefficient alpha was .93, the ICC was .61 [.57–.66] (Skar & Jølle, 2017); the mean score for students was 2.79, and the standard deviation was .86.

Using the equation presented in the section CTT Concepts and Computational Resources, the general

SEM equals 0.53. The interpretation under the CTT approach is that we are 68 % confident (Harville, 1991) that a candidate with a score equaling the mean score of 2.79 has a true score in the range of 2.79 ± 0.53 = 2.26–3.32. Using conventional rounding, this result would imply that a student's true score would be either 2 or 3 which can be considered to be a substantial difference on a five-point scale. If a pass/fail cut score equal to the mean was to be established, 13 of the 25 students would have scores within a confidence interval overlapping that pass score (please refer to column "Obs.Avg" of Table 2).

MFRM Results

An analysis of the 2,400 standardized residuals showed that 109 (4.54%) exceeded |2|. Of those, six (0.25%) were greater than |3|. Therefore, it was concluded that the data fit the model reasonably well. Table 2 presents the observed average score for each student, the scaled score from the MFRM analysis, and the conditional standard error of measurement associated with the latter. The table also contains separation indices and fit statistics for each element. As can be noted, the SEMs were tightly clustered around the mean value of 1.48 (SD = 0.08), and the H_p-index suggests that the measurement allowed for separation of students into 14 distinct statistical stratas. The reliability of the separations was high (R = .99). Although the SEMs showed little variation, there was an expected non-linear relationship between the scaled score and SEM, as is apparent in Figure 1. The scatter plot suggests that the SEMs were lower for students scoring around the mean. If a pass/fail cut score that equaled the mean (≈57) was to be established, again using 68 % confidence, two students would have a confidence interval around their score that overlapped the cut score.

Table 3 displays MFRM results for the rater facet. Applying the same reliability and separation statistics, it becomes evident that the raters also differed substantially. Keeping in mind the counter-intuitive use of reliability for this facet, it can be seen that the reliability of rater separation was .95, indicating that raters differed in severity to a substantial extent. To illustrate the diagnostic value of a MFRM analysis, we have also included descriptive and separation statistics for genres (or "occasion" for G-theory) and rating scales in Tables 4 and 5. For a writing assessment developer it might be useful to know to which extent genres and rating scales differ in difficulty, once disentangled from student ability and rater severity.

G-Theory Results

For the G-theory analysis, the score from each of the five rating scales was summed for each student; thus, the possible scores ranged from 6 to 30. Both raters and genres were treated as random facets, resulting in a $p \times g \times r$ design. The results for the G and D studies are displayed in Tables 6 and 7, respectively. In Table 7, the ANOVA degrees of freedom and mean squares are shown for each source of variation, and the actual variance component estimate and percent of the total

Table 1. Correlations Between Raters (Spearman's rho)

	R1	R2	R3	R4	R5	R6	R7	R8
R1	–	.720	.685	.619	.560	.640	.626	.664
R2		–	.693	.696	.631	.576	.679	.675
R3			–	.635	.599	.607	.644	.617
R4				–	.599	.600	.579	.604
R5					–	.518	.572	.550
R6						–	.559	.646
R7							–	.587
R8								–

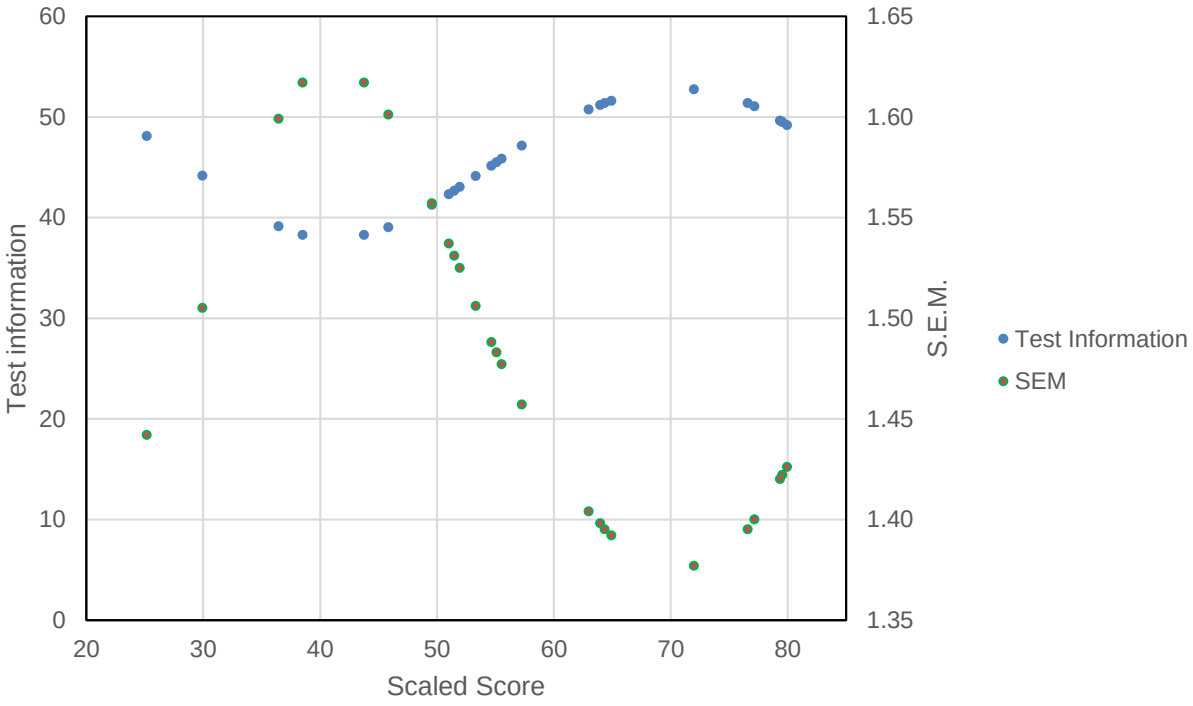
Note. R1 = Rater 1, R2 = Rater 2 and so on. All correlations are statistically significant ($p < .001$).

Table 2. MFRM Results for Students

Student ID	Obs. Avg.	Scaled Score	SEM	Infit
1	2.41	49.57	1.56	.95
2	2.25	45.84	1.60	.84
3	3.10	63.95	1.40	1.22
4	1.36	25.17	1.44	.88
5	3.54	71.98	1.38	1.24
6	2.47	51.01	1.54	.96
7	3.79	76.58	1.40	.98
8	3.82	77.16	1.40	.81
9	3.12	64.34	1.40	.56
10	3.16	64.93	1.39	.79
11	2.64	54.67	1.49	1.12
12	3.97	79.96	1.43	1.04
13	3.95	79.55	1.42	.59
14	2.17	43.77	1.62	.77
15	2.57	53.32	1.51	.78
16	1.96	38.51	1.62	.99
17	1.88	36.44	1.60	.78
18	1.59	29.93	1.51	1.26
19	3.94	79.35	1.42	.81
20	2.76	57.27	1.46	.82
21	2.51	51.95	1.53	.81
22	3.05	62.97	1.40	1.72
23	2.68	55.55	1.48	1.43
24	2.66	55.11	1.48	1.54
25	2.49	51.48	1.53	1.05
Mean	2.79	56.81	1.48	.99
SD	.74	15.45	.08	.28
MSE_p		2.19		
SA_p		236.6		
R_p		.99		
G_p		10.39		
H_p		14.18		

Note. The scaled score was derived using a linear transformation of the logit scores. The transformation enables effective communication to non-technical audiences, since no students will have negative proficiency values.

Figure 1. Test Information and SEM for Students



Note. This plot illustrates the relation between test information and SEM; specifically, the SEM is lower when information is higher and vice versa. TABLE 3

Table 3. Descriptive and Separation Statistics for Raters

Rater ID	Obs. Avg.	Scaled Score	SEM	Infit
R1	2.78	50.26	.83	.94
R2	3.03	45.03	.83	.59
R3	2.68	52.42	.87	1.00
R4	2.63	53.40	.84	.74
R5	2,66	52.84	.84	1.30
R6	2.71	51.79	.84	.98
R7	2.78	50.26	.83	1.36
R8	3.08	44.00	.83	1.04
Mean	2.79	50.00	.83	.99
SD	.17	3.57	.00	.26
MSE_r		.69		
SA_r		12.08		
R_r		.95		
G_r		4.17		
H_r		5.89		

Table 4. Descriptive and Separation Statistics for Genre (Occasion)

Genre / Occasion	Obs. Avg	Scaled Score	SEM	Infit
Expository	2.68	52.27	.42	1.00
Narrative	2.90	47.73	.42	.98
Mean	2.79	50.0	.42	.99
SD	.16	3.21	.00	.01
MSE_o		.17		
SA_o		10.14		
R_o		.98		
G_o		7.65		
H_o		10.53		

Table 5. Descriptive and Separation Statistics for Rating Scales

Rating Scale	Obs. Avg.	Scaled Score	SEM	Infit
Communication	2.91	47.47	.72	1.16
Content	2.80	49.91	.72	1.21
Text Structure	2.77	50.43	.72	.86
Language Use	2.75	50.95	.72	.86
Spelling	2.89	47.89	.72	1.05
Punctuation	2.63	53.36	.73	.82
Mean	2.79	50.00	.72	.99
SD	.10	2.16	.00	.17
MSE_s		.52		
SA_s		4.13		
R_s		.89		
G_s		2.82		
H_s		4.09		

variability are displayed in the rightmost columns. The variance due to persons accounts for the largest percentage, nearly 65% of the total variation. This indicates that, unsurprisingly, students' scores differed substantially. The variances for genres and raters were relatively small (2.4 and 2.5%, respectively). The interaction with the largest variance component is for the persons by genres interaction (P x G), indicating that the relative standings of persons differed somewhat from one genre to the other. Finally, a substantial amount of variance (about 17%) was due to the three-way interaction of person, genre, and rater, and/or other systematic variation not addressed in the study, and/or random noise.

Table 7 shows the values for $\hat{\sigma}^2(\Delta)$ and $\widehat{\Phi}$, the absolute error variance and dependability coefficient, for the original sample sizes $n'_R = n_R = 8$ and $n'_G = n_G = 2$, as well as the projected values of $\hat{\sigma}^2(\Delta)$ and $\widehat{\Phi}$ for some hypothetical sample sizes. While there is

Table 6. G Study for the p × g × r Design

Source	Df	Mean Square	Variance component	Percent of variability
Persons	24	318.9	18.19	64.9
Genres	1	171.6	.68	2.4
Raters	7	53.6	.79	2.8
P x G	24	27.0	2.78	9.9
P x R	168	5.7	.46	1.6
G x R	7	13.0	.33	1.2
P x G x R (Residual)	168	4.8	4.80	17.1

Table 7. D study for the p × g × r Design

	D Studies					
n'_G	1	1	1	2	2	2
n'_R	2	4	8	2	4	8
$\hat{\sigma}^2(\Delta)$	6.65	5.06	4.26	3.64	2.68	2.21
$\hat{\Phi}$	.73	.78	.81	.83	.87	.89

no universally agreed upon "acceptable" level of reliability, Shavelson and Webb (1991) suggest that .80 is "reasonable." The bottom row of Table 7 shows that this level can be nearly reached with one genre and four raters ($\hat{\Phi} = .78$) or slightly exceeded by with two genres and two raters ($\hat{\Phi} = .81$).

Next, G-theory CSEMs were computed for each of the subjects with the original sample sizes using the equation for $\hat{\sigma}(\Delta_p)$ shown in the previous section. However, eight raters may be impractical for many assessment contexts. The ability of the D study to obtain projected reliability estimates under hypothetical samples sizes also extends to the CSEMs. Figure 2 shows the CSEMs for $n'_G = n_G = 2$ genres and n'_R =1, 2, 3 and 8 raters plotted versus the mean over the $n_G * n_R$ =16 scores for each student.

The CSEM for a given person is a function of the variance components for the $R \times G$ G-study based only on that person's data, notated as $\hat{\sigma}^2(R)_p$, $\hat{\sigma}^2(G)_p$, and $\hat{\sigma}^2(RG)_p$. Thus, those subjects with relatively large within-person variance components will have relatively large CSEM values. The largest CSEM (3.64) for $n'_G = n_G = 2$ and $n'_R = 8$ is especially noticeable in the upper right hand section of Figure 2. This is discussed in the next section. As done for previous methods, 68% confidence intervals were created for each student taking their mean score over genres and raters and adding and subtracting (one times) their CSEM. Assuming a cut score was set to the grand mean (17), then six students had intervals containing the cut score.

A Case Study

Student #5 offers an interesting case study. S/he scored well above the mean, both in raw scores, and scaled Rasch scores. Under the G-theory approach, this student had the largest CSEM, but under the MFRM the same student has the lowest. Table 8 provides the raw scores and descriptive statistics for this student. As can be seen, s/he received her or his highest scores on the narrative text. It is also possible to note considerable variation. The large within-in person variation causes MFRM to estimate a small standard error for student #5. However, this student was flagged by the MFRM analysis with significantly high infit and outfit values. These indicate several unexpected results. From a writing proficiency theory perspective, it is odd that student #5 received 2.50 points on spelling in the expository text, and 4.25 in the narrative text, and on punctuation 2.88 and 4.13, respectively. Spelling and other transcription skills are normally automatized and not heavily task sensitive. When the MFRM analysis indicated that student #5 was measured with high precision, as indicated by the low standard error, it is somewhat counter-intuitive: the raw scores suggests that this student had an uneven

Figure 2. G-theory CSEMs Versus Means for $n'_R = 1, 2, 3$ and $n_r = 8$ Raters.

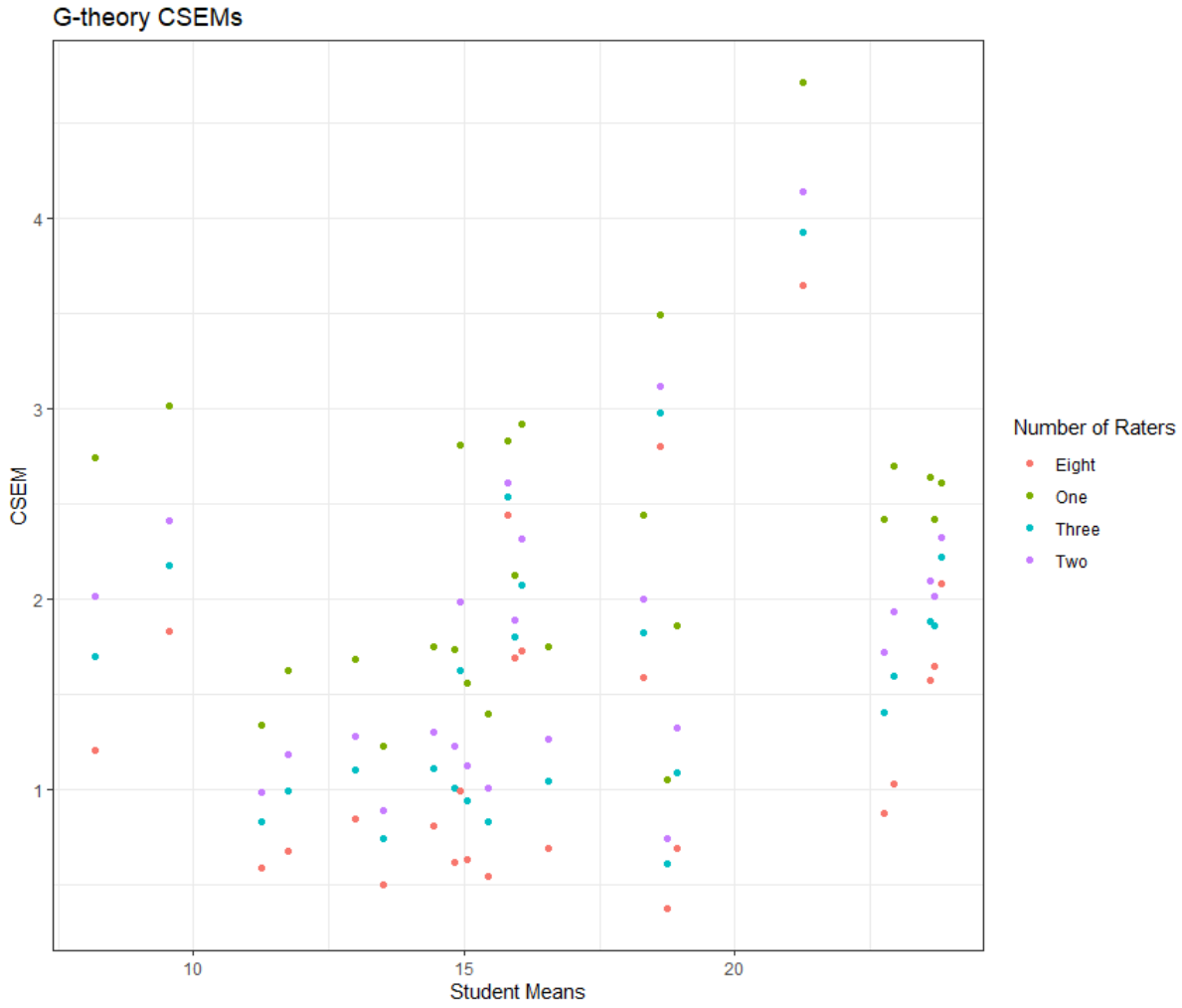


Table 8. Descriptive Statistics for Student #5

	Expository		Narrative		Overall	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Communication	3.38	.92	4.38	.92	3.88	1.02
Content	3.00	.53	4.13	1.13	3.56	1.03
Text Structure	3.13	.64	3.75	.89	3.44	.81
Language Use	2.88	.83	4.13	.64	3.50	.97
Spelling	2.50	.53	4.25	.71	3.38	1.09
Punctuation	2.88	.35	4.13	.83	3.50	.89
Overall	2.96	.68	4.13	.84	3.54	.96

profile, and that it might have been be beneficial to measures his or her skills on more occasions. Inspecting the relationship between SEMs and the infit statistic, Figure 3 suggests that precision, in terms of SEMs, should not be mistaken for model fit. Student with high and low SEMs demonstrated both good and poor fit to the model.

The G-theory CSEM for student #5 for $n'_G = n_G = 2$ and $n'_R = n_R = 8$ is 3.64. The variance components for the within-person $R \times G$ G-study

are displayed in Table 9. The values of $\hat{\sigma}^2(R)_p$ and $\hat{\sigma}^2(G)_p$ are larger than for any other student; this means that the scores for student 5 varied considerably over genres and raters. Since the CSEM is a function of these quantities, student #5 had the largest CSEM. Thus, it is intuitive that G-theory CSEMs are

proportional to the amount of variation of within-student scores. The CSEMs for $n'_G = n_G = 2$ and $n'_R = 8$ and variance components for these persons are shown in Table 9.

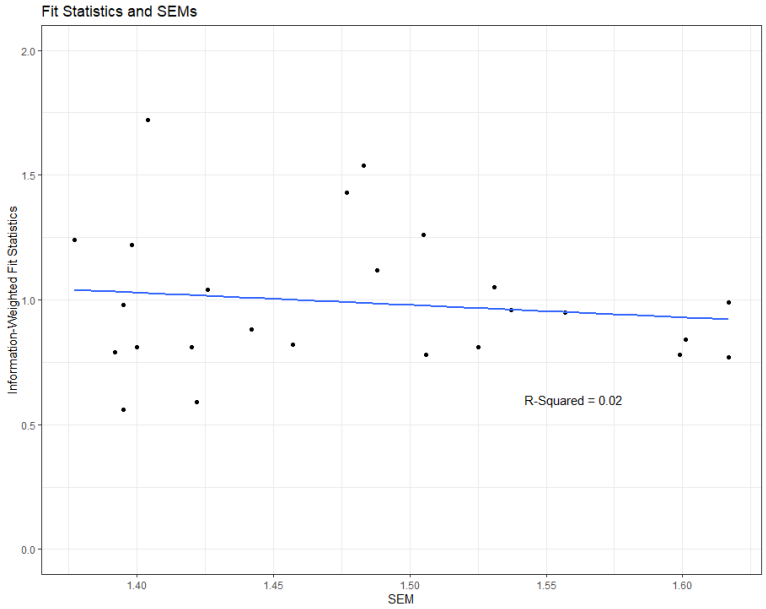
Conclusion

The aim of this tutorial was to present approaches for establishing SEM in writing assessment as well as give detailed accounts of procedures for estimating SEM under a CTT approach, a MFRM approach and a G-theory approach, respectively. We did so by presenting the basic steps in the analysis, and provide examples using real data. The sample size was small and thus the generalizability of the results may be limited; however, we focused on illustrating the process and providing tools for practitioners to analyze their own data.

Table 9. Mean and Within-Person Variance Components for Student #5.

	Mean Score	$\hat{\sigma}^2(R)_p$	$\hat{\sigma}^2(G)_p$	$\hat{\sigma}^2(RG)_p$
Largest CSEM (3.64)	21.25	8.21	23.98	4.00

Figure 3. The Relationship Between Fit Statistics and SEM in MFRM.



Note. SEM cannot easily be predicted by fit statistic; Student with high and low SEMs demonstrated both good and poor fit to the model.

Previous research has investigated merits of the different approaches and compared them in terms of accuracy. This was beyond the scope of this paper, but we have presented some observations of comparative character. First, compared to the others, the CTT approach is fairly uncomplicated and can easily be done using a spreadsheet software like Microsoft Excel®. With that said, compared to MFRM and G-Theory, the general SEM seems to inflate the number of students with large confidence intervals, likely because the traditional reliability estimate is an insufficient proxy for rater effects. The MFRM and G-theory approaches both disentangle effects of facets contributing to variance in scores. Given how variation is treated, G-theory will flag student composite score based on high variation as uncertain, while MFRM will tend to do the opposite. In the concrete case study presented above, student #5 had the largest and smallest SEM, for G-theory and MFRM approaches, respectively. A closer inspection of the raw data suggested a student with somewhat surprising results (such as task-related spelling competence). It is debatable, then, which approach provides the researcher with most useful information. It can seem as the MFRM SEM estimate can be counterintuitive low, but using the infit statistic as a complement may reduce the risk of making faulty interpretations.

References

Aryadoust, V., & Raquel, M. (Eds.). (2019). *Quantitative data analysis for language assessment volume I*. Routledge.

Andrich, D. (2016). Rasch Rating-Scale Model. In *Handbook of item response theory. Volume one*. Models. (pp. 75–94). CRC Press.

Shavelson, R. J., & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch mode*l (3rd ed.). Routledge.

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83–100. https://doi.org/10.1177/0265532214542994

Brennan, R.L. (1998). Raw-score conditional standard error of measurement in generalizability theory. *Applied Psychological Measurement*, *22*(4), 307-331.

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

Brennan, R. L. (2001b). Manual for mGENOVA Version 2.1. Iowa City, IA: Iowa Testing Programs, University of Iowa.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Engelhard, G. (2013). *Invariant Measurement*. Routledge.

Gwet, K. L. (2014). *Handbook of inter-rater reliability*. Advanced Analytics, LLC.

Harvill, L.M. (1991). Standard error of measurement. *Educational Measurement Issues and Practice*, *10*(2), 33-41. https://doi.org/10.1111/j.1745-3992.1991.tb00195.x

Huebner, A. & Lucht, M. (2019) "Generalizability Theory in R," *Practical Assessment, Research, and Evaluation*. *24*(5).https://scholarworks.umass.edu/pare/vol24/iss1/5

Jeffery, J. V., Elf, N., Skar, G. B., & Wilcox, K. C. (2018). Writing development and education standards in cross-national perspective. *Writing & Pedagogy*, *10*(3), 333–370. https://doi.org/10.1558/wap.34587

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). MESA Press.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878. https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2018). A user's guide to FACETS. Rasch-model computer programs. Program manual 3.80.4. Winsteps.com.

Linacre, J. M. (2018). Facets® (version 3.80.4) [Computer Software]. Winsteps.com.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158–180. https://doi.org/10.1177/026553229801500202

McNamara, T. F. (1996). *Measuring second language performance*. Longman.

Moore, C. T. (2016). gtheory: Apply Generalizability Theory with R. R package version 0.1.2. Retrieved from https://CRAN.R-project.org/package=gtheory.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical

Computing, Vienna, Austria. URL http://www.R-project.org/.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. The University of Chicago Press.

Revelle, W. (2019) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.9.12.

Robitzsch, A. (2020). sirt: Supplementary item response theory models. R package version 3.9-4. Retrieved from https://CRAN.R-project.org/package=sirt

Robitzsch, A., Kiefer, T., & Wu M. (2020). TAM: Test analysis modules. R package version 3.5-19, https://CRAN.R-project.org/package=TAM

Schumacker, R. E., & Smith, E. V. (2007). Reliability: A Rasch perspective. *Educational and Psychological Measurement, 67*(3), 394–409. https://doi.org/10.1177/0013164406294776

Skar, G. B. (2017). *The Norwegian National Sample-Based Writing Test 2016: Technical Report*. Nasjonalt senter for skriveopplæring og skriveforsking. http://www.skrivesenteret.no/uploads/files/Skriveproven2017/NSBWT2017.pdf

Skar, G. B., & Jølle, L. (2017). Teachers as raters: Investigation of a long term writing assessment program. *L1 Educational Studies in Language and Literature*, *17*(Open Issue), 1–30. https://doi.org/10.17239/L1ESLL-2017.17.01.06

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4). http://pareonline.net/getvn.asp?v=9&n=4

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261.

Willse, J. T. (2018). CTT: Classical Test Theory Functions. R package version 2.3.3. https://CRAN.R-project.org/package=CTT

Citation:
Huebner, A., & Skar, G. B. (2021). Conditional Standard Error of Measurement: Classical Test Theory, Generalizability Theory and Many-Facet Rasch Measurement with Applications to Writing Assessment. *Practical Assessment, Research & Evaluation*, 26(14). Available online: https://scholarworks.umass.edu/pare/vol26/iss1/14/

Corresponding Author
Alan Huebner
University of Notre Dame

email: alan.huebner.10 [at] nd.edu

Appendix A. Annotated Specification file for FACETS

The following specification was used for the MFRM analysis. A semicolon starts an annotation.

```
Title = [Insert title of analysis here]
Facets = 4; Four facets
Positive = 1; For students, high ability -> high logit. For all other facets, difficulty/harshness -> high logit
Inter-rater = 3; facet 3 is the rater facet
Noncentered= 1; Center the elements all the facets except facet 1. This established the zero point (local origin) of the measurement frame-of-reference
Umean = 50, 10, 3 ; set mean to 50, logit distance to 10 and report three decimals
Model= ?,?,?,?,R5 ; The Andrich rating scale model for judges, persons, tasks and rating scales. Highest possible mark on a rating scale was 5 (all other values are automatically treated as missing).
*
Labels=
1, Student
1-561; the 25 students were numbered 1–561. Unobserved elements are excluded from the analysis.
*
2, Genre
1584=Expository; in the data file, each genre had a code
1585=Narrative
*
3, Rater
9=R1
10=R2
13=R3
16=R4
17=R5
22=R6
38=R7
45=R8
*
4, Rating Scale
1=Communication
2=Content
3=Text Structure
4=Language Use
5=Spelling
6=Punctuation
*
Data =
1,1584,9,1-6,2,2,2,2,3,2; This is the first data string. It equals student #1, task 1584 (expository), rater #1, rating scales 1–6, and scores 2, 2, 2, 2, 3 and 2 on each of the six scales.
```

Instructions for calculation test information for element

This guide was provided by Linacre (personal communication):

- In FACETS, click “Output Files” and choose “Residuals/Responses files”.
- Choose “Output to Excel”.
- In Excel, sort the file on the elements you wish to compute test information for.
- Sum the column “Var” for an element.

Appendix B: G-theory CSEM code example

We present reproducible examples of computing CSEMs for G-theory using the code provided in Appendix C. Both examples use the data from Table 3.2 of Brennan (2001), which is publicly available in the *gtheory* package. We present examples for one- and two-facet G-theory designs, in turn. The functions in Appendix C should be copied and pasted into an R script, the code provided below should also be pasted in the same script. Note, this tutorial covers obtaining CSEMs only; for information on obtaining basic G and D study results for G-theory, see Huebner and Lucht (2019).

One-facet Design

We load the *gtheory* package as well as the *dplyr* package, which is used by the functions provided. Then, the data for Table 3.2 of Brennan is loaded from the gtheory package:

```
library(gtheory)
library(dplyr)
data("Brennan.3.2")
```

The data is originally from a nested two-facet design, but we recast it as a one-facet crossed $p \times t$ design, with $n_p = 10$ subjects performing the same $n_t = 12$ tasks. The resulting data is named dat1:

```
dat1 <- Brennan.3.2
dat1$Task <- rep(1:10, times = 12)
```

The function calcGtheory1FacetCSEM() has the following arguments:

```
calcGtheory1FacetCSEM(Person, Facet, Score, nf_prime = NULL)
```

The vectors for person, facet and score are the first three arguments. The fourth argument is the number of levels for the facet, or n_t' for facet t. If a value is not specified, the actual number of instances $n_t' = n_t$ is used. Or, the user can specify a value of n_t' that is not equal to n_t. For example, running the following code will return CSEMs based on the actual number of tasks, $n_t' = n_t = 12$:

```
calcGtheory1FacetCSEM(dat1$Person, dat1$Task, dat1$Score)
```

Alternately, we can obtain CSEMs based on, for example, $n_t' = 8$ items:

```
calcGtheory1FacetCSEM(dat1$Person, dat1$Task, dat1$Score, 8)
```

Accordingly, the CSEM values for the second run with $n_t' = 8$ are larger than for the first run with $n_t' = n_t = 12$.

Two-facet Design

For the two-facet design, the rater facet from Table 3.2 is incorporated into the analysis, and the data is recast as a crossed two-facet $p \times t \times r$ design, with $n_p = 10$ subjects, $n_t = 3$ tasks, and $n_r = 4$ raters using the following code:

```
data("Brennan.3.2")
dat2 <- Brennan.3.2
dat2$Rater <- rep(c(1:4), each = 10, times = 3)
```

The function calcGtheory2FacetCSEM() has the following arguments:

```
calcGtheory2FacetCSEM(Person, Facet1, Facet2, Score, nf1_prime = NULL, nf2_prime = NULL)
```

The vectors for Person, the two facets, and score are the first four arguments. The fifth and sixth arguments are the numbers of levels for the first and second facets, i.e. n_t' and n_r' respectively. If values are not specified, the actual sample sizes $n_t' = n_t$ and $n_r' = n_r$ are used. For example, exectuting the following code will return CSEMs based on the actual numbers of tasks and raters, $n_t' = n_t = 3$ and $n_r' = n_r = 4$, respectively:

```
calcGtheory2FacetCSEM(dat2$Person, dat2$Task, dat2$Rater, dat2$Score)
```

Or, for example, we can obtain the CSEMs when there are two tasks and two raters:

```
calcGtheory2FacetCSEM(dat2$Person, dat2$Task, dat2$Rater, dat2$Score, 2, 2)
```

Appendix C: R Functions for Computing G-theory Conditional Standard Errors

```
#Function to compute CSEM for 1-facet G-theory designs.

calcGtheory1FacetCSEM <- function(Person, Facet, Score, nf_prime = NULL){
  #Obtain sample sizes
  np <- length(unique(Person))
  df1 <- data.frame(Person, Facet, Score)
  if (is.null(nf_prime)) nf <- length(unique(Facet)) else nf <- nf_prime
  var_persons <- df1 %>% group_by(Person) %>%
                      summarise(VarPers = var(Score)) %>% pull(VarPers)
  AbsCondCSEM <- sqrt(var_persons/nf)
  CSEM <- data.frame(unique(Person), AbsCondCSEM)
  names(CSEM) <- c("Person", "AbsCondSEM")
  return(CSEM)
}

#Function to compute CSEM for 2-facet crossed G-theory designs.

calcGtheory2FacetCSEM <- function(Person, Facet1, Facet2, Score, nf1_prime = NULL, nf2_prime = NULL){
  #Obtain sample sizes
  np <- length(unique(Person))
  if (is.null(nf1_prime)) nf1 <- length(unique(Facet1)) else nf1 <- nf1_prime
  if (is.null(nf2_prime)) nf2 <- length(unique(Facet2)) else nf2 <- nf2_prime
  data_csem <- data.frame(Person, Facet1, Facet2, Score)
  #Formula for within person variation
  formula_i <- Score ~ (1|Facet1) + (1|Facet2)
  #Save results for conditional absolute SEM
  AbsCondSEM <- numeric(np)
  #Loop through subjects; compute conditional absolute SEMs for each
  for (i in 1:np) {
    #Get data for student i
    dat_i <- data_csem[data_csem$Person == unique(data_csem$Person)[i],]
    #perform G Study and extract variance components
    gstud_i <- gstudy(dat_i, formula_i)
    var_i <- gstud_i$components[,2]
    #Compute conditional abs SEM for subject i and save
    AbsSEM_i <- sqrt(var_i[which(gstud_i$components[,1] == "Facet1")]/nf1 +
                  var_i[which(gstud_i$components[,1] == "Facet2")]/nf2 +
                  var_i[3]/(nf1*nf2))
    AbsCondSEM[i] <- AbsSEM_i
  }
  CSEM <- data.frame(unique(data_csem$Person), AbsCondSEM)
  names(CSEM) <- c("Person", "AbsCondSEM")
  return(CSEM)
}
```