2012; 34: 960–992 MEDICAL TEACHER

AMEE GUIDE

Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68

RALPH BLOCH & GEOFFREY NORMAN

McMaster University, Canada

Abstract

Background: Generalizability theory (G theory) is a statistical method to analyze the results of psychometric tests, such as tests of performance like the Objective Structured Clinical Examination, written or computer-based knowledge tests, rating scales, or self-assessment and personality tests. It is a generalization of classical reliability theory, which examines the relative contribution of the primary variable of interest, the performance of subjects, compared to error variance. In G theory, various sources of error contributing to the inaccuracy of measurement are explored. G theory is a valuable tool in judging the methodological quality of an assessment method and improving its precision.

Aim: Starting from basic statistical principles, we gradually develop and explain the method. We introduce tools to perform generalizability analysis, and illustrate the use of generalizability analysis with a series of common, practical examples in educational practice.

Conclusion: We realize that statistics and mathematics can be either boring or fearsome to many physicians and educators, yet we believe that some foundations are necessary for a better understanding of generalizability analysis. Consequently, we have tried, wherever possible, to keep the use of equations to a minimum and to use a conversational and slightly "off-serious" style.

Introduction

Although we wrote this monograph primarily for Members of the Association for Medical Education in Europe (AMEE), it could be of interest to any serious medical educator, in fact, any educator who is involved with the development and administration of assessment procedures.

Society, appropriately, is concerned with the professional competency of physicians, yet it lacks the prerequisite ability to supervise it. Consequently, it has delegated the responsibility for quality assurance to the professional colleges and medical schools. These, in turn, have built up a veritable "assessment industry". But, who assesses the assessment? Thus, we have all gradually become increasingly conscious of the need for quality assurance of high stakes assessment. One of the most powerful tools to explore the value of methods to evaluate knowledge, skills and, possibly, attitudes, is generalizability theory or as it is more commonly known, G theory.

Yet, for many of us, G theory is still a black art. Basically, it (G theory) explores the fundamental question: to what extent can we extrapolate the results achieved on a limited sample of test tasks, measured under unique test conditions to a universe of tasks and conditions, from which the specific test set has been drawn more or less arbitrarily.

The literature on G theory is no easy fare, nor do tools for G theory data processing abound. Some four years ago, we started to develop a computer program – G_String – to give evaluation practitioners a tool to analyze their data using G theory. G_String wraps around a command line program, performing the core calculations, called urGENOVA (University of Iowa), written by Robert L. Brennan, one of the leading experts in the field. The name of our software, "G_String", has raised more than one eyebrow and academic firewall concerned with propriety. In fact, the semantics are quite innocent: "G" stands for generalizability, and "Strings" are lexical sequences of symbols (letters) which the program

Practice points

- Testing knowledge and performance is a measurement.
- Measurements provide a mixture of true data (signal) and confounders (noise).
- Statistical methods like G theory allow us to separate noise from signal, identify sources of noise, and devise ways to reduce their contribution to the final results.
- G theory is a powerful method to achieve this goal.
- G theory is an extension of the two-factor, random-model ANOVA.
- G theory, like any analytical tool, is only useful if it is accompanied by careful experimental design, planning, and analysis.

Correspondence: R. Bloch, Faculty of Health Sciences, McMaster University, 1280 Main Street West, Hamilton, ON, L8S 4K1, Canada. Tel: 905-525-9140 ext. 23114; fax: 905-572-7099; email: blochr@mcmaster.ca

ISSN 0142–159X print/ISSN 1466–187X online/12/110960–33 © 2012 Informa UK Ltd.
DOI: 10.3109/0142159X.2012.703791

parses and analyses in order to instruct urGENOVA on how to perform its calculations and what to do with the results of these calculations. G_String is freely available and can be downloaded free from: http://fhsperd.mcmaster.ca/g_string/index.html at McMaster University.

G_String is not the only computer program of its kind. Others are available as well, but it is the program we know best. This is why we have used it as the backbone in this monograph.

We are not attempting to turn you, the reader, into a hard core statistician. Rather, we provide an overview of the theory underlying G theory analysis. For those of you who want to know more, there are books that offer a relatively painless introduction to statistics (Streiner & Norman 2008a). For the more serious among the readers are typical classics (Winer et al. 1991). For those, who live on the web, there is a wonderful Statistics portal by NIST (National Institute of Standards and Technology 2003). Finally, for the more pragmatic amongst the readers, almost any technical term is explained in Wikipedia and can be easily found on Google (www.google.com).

In "Signal and noise" section, we introduce the concept of statistical noise as well as "signal-to-noise ratio". "Two-way ANOVA" section provides an example of a classical analysis of reliability using simple two-way analysis of variance (ANOVA). In these first two sections, we emphasize a mathematical approach. While the reader does not necessarily have to trudge through the formulae in order to be able to apply the topics of subsequent sections, it would surely add to a deeper understanding as it demonstrates how generalizability theory is an extension of classical test theory (CTT) using mathematical statistics.

In "Beyond CTT" section, we take a more detailed look at sources of experimental noise and extend the theory to multiple factors. We also introduce the basic terminology of G theory. We abandon the formal mathematical derivation at this point, and refer the interested reader to appropriate sources. In "Research design" section, we demonstrate how these concepts mesh with research design in evaluation. "Designing your G study" provides a general approach to designing G studies, i.e., studies to analyze the properties of a given test. "Computing G coefficients" section explains how G coefficients are calculated. This is not essential; G_String performs these calculations, but it may help achieve better understanding of how the coefficients have been calculated by the software. "G theory software" section discusses the available G theory software and explains G_String in more detail. "Worked example" section provides a medley of worked examples to illustrate the possibilities.

We have specifically added Appendices A–D which focus entirely on the use of G_String. Appendix A illustrates how one goes about analyzing a dataset with G_String. Appendix B is more technical, it describes the data requirements of G_String. Appendix C explains the program output and how to interpret it. Appendix D, finally, lists possible error messages of G_String and explains their significance.

Since it is easily downloadable and free, we would recommend that G_String is used at the same time as reading and using this Guide.

Signal and noise

Outline

- Why G theory?
- The basic idea of signal and noise
- Ratio of S/(S+N)

Why G theory?

You are sitting across the table from the Minister of Health and her coterie (a small exclusive group of people who share the same interests). She seems rather tense. No wonder, she has been getting a lot of criticism lately for the poor quality of health care services. She is looking for a quick and media-effective way out. And you, lucky fellow, you are it.

"*It's the low quality of our doctors*" says she. "*We have to introduce better quality control!*" With your political knowledge you inquire: "*What did you have in mind?*" She snuffs and replies: "*Don't you see, half the physicians provide below average care? We have to stop this!*"

There is your mission. How do you determine quality of care, and how do you remove physicians who are lacking? While quality-of-care indicators for specific conditions exist, it is neither practical nor economically feasible to accurately monitor the overall quality of care for every physician; and using such indicators in a coercive manner would be politically out of the question. We are thus left with assessing future physicians at the transition from education to training and from training to practice. In a limited way it may be possible to identify low-end-outliers and, hopefully, prevent them from entering practice.

Assessing the qualities of future physicians remains a substitute for controlling actual quality of care. The challenge, thus, becomes finding appropriate and cost-effective measures, applying them efficiently, and demonstrating that they adequately measure competence, and predict future quality of care. But that raises another conundrum: can physician competence be measured on a single scale, or does it rather consist of a basket of individual, independent competencies which are required in differing proportions for different specialties and practice conditions? Finally, if competence cannot be reduced to a single scale, how do you define a threshold below which progression toward independent practice becomes impossible?

Ideally, we could plug candidates into a calibrated black box that provides us with a digital readout of their competence. If it had a chute for rejects, we would not even have to pick up the pieces. But, of course it just is not that simple. A battery of multiple choice questions may be adequate for testing factual knowledge and maybe some reasoning, but it does not suffice for testing professional skills which, unfortunately, constitute a major portion of physician competence. Skills manifest themselves as purposeful behaviors appropriate to specific circumstances, and need to be judged by a knowledgeable observer.

The magnitude of the skills inventory required for the practice of medicine makes comprehensive testing impossible. Practical and economic considerations severely limit the

sample size of requisite skills that can be tested in an exam. Thus, from the outset a clinical skills examination can only provide a crude estimate of a candidate's skill level. The smaller the test set the cruder the estimate. Since the skills score requires human judgment, extraneous factors will affect the result. One of the authors' centenarian mother reminisces that she made it a habit to always wear a short red dress for oral exams; she knew about observer subjectivity!

For any large test involving many candidates and many test situations (cases, stations, orals), different candidates may meet different raters and different standardized patients even if efforts are made to control the clinical case presentation. There are multiple sources of variability. We have already mentioned that "medical competence" does not fit on a simple, one-dimensional scale, but takes on different values, depending on the specific skill and situation under which it is estimated. There are other sources of error. A candidate may function better on one day than another; she may do better in the morning than in the afternoon. There are also some structural considerations. Maybe you are testing candidates from different schools. True, your primary interest is the competence of individual candidates, but you may also be interested in the teaching effectiveness of the individual schools – yet another source of variability. And so it goes. It is the bane of social sciences that potential sources of variability are almost unlimited. You are dealing with observational data, so randomizing extraneous sources of statistical variability is all but impossible.

And there lies the power of generalizability theory; it allows you to estimate the contributions of different sources of variance, as long as you can group your individual measurements appropriately, and you are able to estimate all individual data items on an identical numerical scale. The variability of primary interest to you is the differences in estimated competence between the different candidates – we call this "the signal". Any other source of variability constitutes statistical "noise". In designing and refining your testing procedure, you pay attention to the various sources of noise in order to minimize them.

The basic idea of signal and noise

The related concepts of signal and noise arose originally during the development of RADAR in WW II Britain. Signal was what you were interested in; it meant "enemy aircraft approaching". Noise meant distraction; it might have masked a true signal or falsely sent Royal Air Force fighters on a wild Messerschmitt chase. With the progress of signal theory, this qualitative approach was soon replaced by much more quantitative and mathematical methods.

But signal theory is just the special case where statistics is applied to time series – commonly of continuous electrical voltage or current. Thus it appears quite natural that such descriptive terms like "signal" and "noise" became generalized. However, instead of defining "signal" as the total presence or absence of signal amplitude, it became focused on a defined change in signal amplitude. Before we can visualize "signal-to-noise", we have to familiarize ourselves more quantitatively with the nature of noise.

You may be a statistical maven, or like the rest of us, you might, at times, become confused by the apparent leaps of faith, statisticians employ. Much of this confusion stems from the fact that statistics really consists of two distinct concepts in one wrapper. First, statistics deals with concrete, real, finite data items, the kind of data items you collect and explore in your daily work. Statistics tells you how to manipulate these data. The second concept – much more abstract – deals with idealized, infinite datasets. What ties those two concepts together is a theory of how you infer properties of the infinite set from the finite data items.

Let us assume for the moment that we always employ the same limited test set for our competence exam: the 10 most common clinical situations. In this case, test results do not allow us to generalize to the general competence level of candidates; only that they can handle the 10 most common clinical situations. If we want to be able to generalize to a larger universe of clinical skills, we have to draw our test set randomly from that universe. But to do so, we have to have some way of estimating the extent to which the specific characteristics of the small set of 10 situations may lead to error in the generalization. To return to our first example, the extent to which the specific rater, standardized patient, or case may lead to a biased estimate, or the effect of the time of day, the school where they are being tested, or any number of other variables.

Maybe we have got ahead of ourselves. We need to begin by quantitatively examining the issue of measurement errors in a simpler situation. Let us start over with just about the simplest example possible: here is a string and a ruler. Your task is to measure the length of the string. We will assume that you are lucky, you own the absolutely accurate reference ruler that sets the standard for all the other rulers in the world. You realize that a string is an iffy thing. Its length depends, among others, on the tension applied. So, you standardize that by clamping at a uniform tension. Done! Your ruler shows the string to be 101.0 mm long. But your buddy tries the ruler and gets 100.5 mm. Clearly, a single measurement does not mean much, no matter how careful you are.[1] There still remain any number of factors that might affect your final result, which you have not taken into consideration. You may want to standardize temperature and humidity, yet still, repeated measurements result in a range of numbers. There is no way around it, you have to repeat the measurement process a number of times – say "N" times.

The finite, concrete dataset consists of the results of your repeated measurements. The idealized, infinite dataset consists of a continuum of possible true length values of the string. The problem is you really do not know the string's true length. A central tenet of statistics postulates that you will get the "true" value only if you measure the object (string) an infinite number of times. Problem is, that may take you forever, and you have better things to do.

We have been holding back as long as we could, but we just have to introduce some mathematics.

Consider the formula:

$$X_i = \mu + \varepsilon_i \quad \text{with} \quad i = 1, \ldots, N,$$

where X_i stands for the concrete ith measurement of string length, lower case μ stands for the unknown, true value, a specific choice from the idealized, infinite set of possible values, and the unknown ith error term ε_i has been drawn from the idealized, infinite universe error set. Statisticians tend to use capital letters for finite sets and lower case letters for infinite sets. If your measurement errors are truly random, the mean of the error terms tends toward zero as the number N of repeated measurements grows. As a result, the mean value of X_i, namely

$$\langle \mu \rangle = \bar{X} = \frac{1}{n}\sum_{i=1}^{N} X_i$$

will tend toward the elusive "true" value μ with growing N. But what do we mean by "tend toward?" It means, on average, the absolute difference between the true value and the mean of measured values will get smaller and smaller, but cannot ever be expected to become zero for any finite number of N. In other words, it is the best estimate for the true value of μ under a given number of repeated measurements. Statisticians will say: the mean is the best, unbiased estimator of the true value.

Now let us look at the error terms. For each measurement X_i, the error term is just $\langle \varepsilon_i \rangle = X_i - \bar{X}$, where the pointed brackets indicate "estimator." But the individual ε_i estimates do not interest us. We need some kind of aggregate for the error terms. As we have seen above, the estimate of the mean error term is zero, because the error terms are assumed to have a symmetrical distribution around the mean.

To stop the positives and negatives canceling each other, we estimate the squared error (variance) by calculating the mean-squared error of our measurements:

$$\langle \varepsilon_i^2 \rangle = \frac{1}{N-1}\sum_{i=1}^{N}\{X_i - \bar{X}\}^2 = \frac{\sum_{i=1}^{N} X_i^2 - N \times \bar{X}^2}{N-1}$$
$$\equiv \frac{\mathrm{SS}}{\mathrm{df}} \equiv \mathrm{MS}(X) \equiv \mathrm{var}(X) \equiv \sigma^2.$$

Assuming a normal distribution for the individual length measurements, MS is an unbiased estimate for ε^2 or variance. "SS" stands for "sum of squares", "MS" for "mean square (difference)" and "df"(equal to $N-1$ in this case) for "degrees of freedom" and var(X) is the variance of X. In "two-way ANOVA" section, we will employ this terminology again.

Let us say a word about the term "degrees of freedom". There are various formal definitions in the literature. The simplest one found in the Wikipedia is: "*the number of values in the final calculation of a statistic that are free to vary*". In our case, we have N measured values for X_i. But the estimated value of μ is already fixed. In other words, it is the best estimate for the true value of μ. So you only have $N-1$ values left to vary, that is: $\mathrm{df} = N-1$.

If we are attempting to detect a specific, consistent change in the variable X, namely ΔX, we can define the so-called "effect size" ES as

$$\mathrm{ES} = \frac{\Delta X}{\sigma}.$$

As a carryover from signal theory, it is common to consider the power ratio or the square of the effect size and call it "signal-to-noise ratio" or SNR:

$$e^2 = \frac{\Delta X^2}{\overline{X_i^2} - \bar{X}^2} \equiv \frac{\Delta X^2}{\sigma^2} \equiv \mathrm{SNR}.$$

Another common indicator for the quality of a measurement is the so-called "intra-class correlation" (ICC) (Fisher 1925):

$$\mathrm{ICC} = \frac{\mathrm{SNR}}{1+\mathrm{SNR}}.$$

It is the ratio of the variance attributed to the variable of interest to the total variance, so expresses the proportion in the variance of the observed scores that is due to true differences in the variable of interest.

Before we put this section to bed, let us consider the RADAR example once more. The larger the signal relative to the noise, or alternatively, the larger the proportion of the total amplitude that is due to signal, the better your chances of finding the bad guys. So this means that the closer the ICC is to 1, the better the ability to detect the signal. An ICC of 1 says that it is all signal; you have managed to create a perfect, noiseless detection system. An ICC of 0 says, conversely that you are unable to find any signal.

We conclude "Signal and noise" section by summarizing the terms used in estimating the mean and variance of a random variable:

Term used	Symbol	Formula	Example string ($N=10$)
Mean	$\bar{X}$	$\frac{1}{N}\sum_{i=1}^{N} X_i$	10.01 cm
Sum of squares	SS	$\sum_{i=1}^{N} X_i^2 - N \times \bar{X}^2$	$=0.09\,\mathrm{cm}^2$
Mean square (variance)	MS	$\frac{\mathrm{SS}}{\mathrm{df}}$	
Degrees of freedom	df	$N-1$	
Signal-to-noise ratio	SNR	$\frac{\Delta X^2}{\overline{X_i^2} - \bar{X}^2}$	
Intra-class correlation coefficient	ICC	$\frac{\mathrm{SNR}}{1+\mathrm{SNR}}$	

Summary

In this section, we have introduced the scope of G theory as a tool to examine a variety of confounders in the systematic assessment of knowledge and competence. We have also laid a basis for quantitative understanding of signal and noise contained in empiric data.

Two-way ANOVA

Outline

- The simple linear model
- Two way ANOVA
- How to compute
- Numeric example

Now that we have conquered some fundamental statistical principles and measuring the length of a string bears no horror anymore, we can graduate to a simple multiple choice test. Things are kind of the same, but different from "signal and

noise" section. This section, at first glance may be still a bit mathematically loaded, but it is not our purpose to turn you into an expert statistician. Rather, we are trying to illustrate the lineage of G theory, so you can more easily appreciate its strengths and limitations.

It is self-evident that easy tests result in higher scores than difficult tests, and competent students receive higher scores than incompetent ones. The nice thing about self-evident truths is that we can hang on to them, no matter what. However, it is not quite clear how test easiness and student competence interact. The simplest assumption is that they act additively. Let us try that and see what we get. We expect, therefore, that the test score of student "i" on test item "k" looks somewhat like:

$$\text{Expected score}_{i,k} = \text{grand mean} + \text{competence}_i + \text{easiness}_k + \text{fudge factor}_{i,k}.$$

The reason that we need a fudge factor is twofold. Firstly, we have already mentioned that we have made the assumption of additivity somewhat arbitrarily. It is quite likely that some interaction between specific test questions and specific students exists. Secondly, we have lost all kinds of additional information: the fly that buzzed around student "i" while he was dealing with item "k". And there are many other potential sources of noise, too numerous to mention.

Unfortunately, no self-respecting journal would accept an article containing fudge factors. Let us, therefore, make the above formula more scientifically respectable by renaming "fudge factor" to "error term" ε.

$$es_{i,k} = \mu + c_i + e_k + \varepsilon_{i,k}.$$

There you have it. Does not that look scientific? Instead of having just a single factor, the string length as in "signal and noise" section, we now have two: competence (the factor we are really interested in) and easiness. Our task thus becomes to solve a whole swatch of similar linear equations:

$$X_{i,k} = \langle\mu\rangle + \langle c_i\rangle + \langle e_i\rangle + \langle\varepsilon_{i,k}\rangle,$$

where $X_{i,k}$ are the actual scores.

Solving hundreds of simultaneous equations – child's play!

$$\langle\mu\rangle = \frac{1}{N_s \times N_f}\sum_{i=1}^{N_s}\sum_{k=1}^{N_f} X_{i.} \equiv \bar{X} \equiv X_{..},$$

$$\langle c_i\rangle = \frac{1}{N_f} \times \sum_{k=1}^{N_f} X_{i,k} - \bar{X} \equiv X_{i.} - X_{..},$$

$$\langle e_k\rangle = \frac{1}{N_s}\sum_{i=1}^{N_s} X_{i,k} - \bar{X} \equiv X_{.k} - X_{..},$$

where N_s and N_f stand for the number of students and test items, respectively. Now that we have unbiased estimates for the grand mean, for competences, and for easiness, our task is almost completed. The only thing left is this pesky problem of noise.

As in "signal and noise" section, we are not really interested in calculating the individual error term for each measured score. Rather, we want to calculate a compound measure of errors attributable to the two sources of error: students and questions. We will first calculate the total sum of squares:

$$\text{SS}_{\text{total}} = \sum_{i=1}^{N_s}\sum_{k=1}^{N_f}\left(X_{i,k} - X_{..}\right)^2 = \sum_{i=1}^{N_s}\sum_{k=1}^{N_f} X_{i,k}^2 - N_s \times N_f \times X_{..}^2.$$

For the sum of squares attributable to students we get:

$$\text{SS}(s) = N_f \times \sum_{i=1}^{N_s}(X_{i.} - X_{..})^2 = N_f \times \left\{\sum_{i=1}^{N_s} X_{i.}^2 - N_s \times X_{..}^2\right\}.$$

The factor N_f looks confusing at first. But remember, we actually have to perform the sum over all the scores, so we do need $N_s \times N_f$ terms. Similarly, for the sum of squares attributable to the questions, we get:

$$\text{SS}(f) = N_s \times \sum_{k=1}^{N_f}(X_{.k} - X_{..})^2 = N_s \times \left\{\sum_{k=1}^{N_f} X_{.k}^2 - N_f \times X_{..}^2\right\}.$$

Since the error sum of squares is equal to the individual error sums for students plus that of the questions plus the residual error, we can calculate the residual error sum as

$$\text{SS(error)} = \text{SS}_{\text{total}} - \text{SS(s)} - \text{SS}(f).$$

Getting the mean squares is almost trivial in comparison:

$$\text{MS}(s) = \frac{\text{SS}(s)}{N_s - 1}$$

$$\text{MS}(f) = \frac{\text{SS}(f)}{N_f - 1}$$

$$\text{MS(error)} = \frac{\text{SS(error)}}{(N_s - 1) \times (N_f - 1)}$$

We are almost home. Just take a deep breath! Remember, in the "signal and noise" section the variance of string length was simply equal MS (length). But here we have two factors – students and questions, so we have to do a "Two way ANOVA." Estimating the variance components $\sigma^2(s)$, $\sigma^2(f)$, and σ^2(residual) is not computationally difficult.

While the MS(error) gives us the estimate for the variance component due to residual error, the other two mean squares are slightly more complicated. For example, MS(s) contains not only the variance inherent in the differing competence of students (once for each form) but also the actual error or residual variance.

$$\text{MS(error)} = \sigma^2(\text{residual}),$$

$$\text{MS}(d) = \sigma^2(\text{residual}) + N_f \times \sigma^2(s),$$

$$\text{MS}(f) = \sigma^2(\text{residual}) + N_s \times \sigma^2(f).$$

We have finally arrived. We can now calculate estimates for the different variance components:

$$\sigma^2(\text{residual}) = \text{MS(error)},$$

$$\sigma^2(s) = \frac{\text{MS}(s) - \text{MS(error)}}{N_f}.$$

$$\sigma^2(f) = \frac{\text{MS}(f) - \text{MS(error)}}{N_s}.$$

Knowing the estimated variance components is nice. But what do you do with it? The variance component attributable to the exam questions does not interest us too much at this point. What we would like to know is: how much of the observed variance of students' competence is due to their

actual competence, and how much is due to measurement error? Easy:

$$\rho^2 = \frac{\sigma^2(s)}{\sigma^2(s) + \sigma^2(\text{residual})} \equiv \frac{\text{MS}(s) - \text{MS}(\text{error})}{\text{MS}(s) + \left(N_f - 1\right) \times \text{MS}(\text{error})}.$$

The variable ρ^2, sometimes called $E\rho^2$ or "ICC" is a dimensionless number between 0.0 and 1.0. An ICC of 0 is bad, very bad. It gives us nothing but noise. An ICC of 1.0 is good, very good. You have reached the Nirvana of an error free test. Please call the authors immediately and tell us how you did it. Anything between those limits is realistic. For high stake exams, you would like ICC to be well over 0.8. For formative evaluation, values of 0.6–0.7 are more typical.

The ICC is closely related to the SNR mentioned earlier:

$$\rho^2 = \frac{\text{SNR}}{\text{SNR} + 1}.$$

To give more realism to the (synthetic) multiple choice test, we will illustrate how the numerical results could look like in more detail:

Effect	*N*	df	Mean	SS	MS	σ^2
Total	600	1	72.3	4189971.06		
Student	60	59	0	2348974.08	39813.12	4328.64
Form	10	9	0	1386800.82	154088.98	2597.18
Residual		531	0	454196.16	855.36	855.36

The ICC is then $4328/(4328+855) = 0.835$. The value is actually quite reasonable. Only 16.5% of the observed variance of student scores is due to error.

For this fully balanced, crossed design with one facet, the reliability coefficient tells you all you need to know. The two-way ANOVA is a powerful tool to calculate the variance components attributable to students, questions, and residual error, and to determine reliability as ICC.

But in the real world, there may be many more facets. G theory provides you with a generalized tool to estimate variance components under a variety of experimental conditions. Systematically analyzing the variance contribution of various facets allows you to optimize your assessment tool. Think of G theory as reliability on steroids, but more about that in "Beyond CTT" section.

Summary

In this section, we have expanded the discussion of signal and noise, and introduced a fundamental statistical technique: the two-factor random-model ANOVA.

Beyond CTT

Outline

- Classical test theory
- What is wrong with classical test theory
- Basic concepts
  - The "Object of measurement"
  - Facets of generalization
  - Stratification facets

The example we have used to date, looking at how 60 students performed in a simple multiple choice test involving 10 questions, yielding 600 numbers, was our first foray into the arcane discipline called "psychometrics". We calculated an ICC, which represented the proportion of the total variance in the numbers that was due to real differences between students. This is called a "reliability coefficient" and measures the ability of the measuring instrument, the multiple choice test, to discriminate between high and low scoring students. To remind you of the formula:

$$\text{ICC} = \frac{\text{var(students)}}{\text{var(students)} + \text{var(error)}}.$$

Note a few properties of this ratio. First, as we mentioned in "two-way ANOVA" section, it is a number between 0 and 1, with 0 indicating that all the variability in scores is due to error or noise alone, and 1 indicating that all the variability is due to real differences between students. But note that the signal is not simply a number; it is actually a *variance*, measuring how much difference arose between high and low scoring students. If there is no difference between students – everyone got 9/10 – then the reliability is zero, by definition. So reliability is not the same as agreement, since if everyone gets 9/10 there is 100% agreement. Instead it is a measure of discrimination – in the technical, not sociological, sense; the ability of the instrument to distinguish between high and low scorers. One consequence of this formula is that the more homogeneous the sample (and population) the lower the reliability, all other things being equal. So that reliability is only meaningful when you specify the variance of the population you are applying it to.

Now let us talk a bit more about the denominator. As we have seen, it is made up of two variances – variance due to differences among students, also called true variance and a second term that we have called "error". Actually it is a bit more complicated than that. There are two contributors to the error term; the "item × student" interaction – the extent to which some students do well on some items and poorly on others, and others get different items right or wrong, and also random error – what might result if we gave exactly the same items to the students again (assuming they cannot remember their prior response). In fact, because we only have one observation in the individual cell defined by each student and each item we cannot separate the two terms. So as a general rule, we always consider that the highest order interaction is the same as the random error term.

There is also potentially a third source of error – systematic differences between items. This would show up in the analysis as the "main effect" of item and had a variance of 2597 in the previous example. However, in the circumstances we described, where all 60 students got the same test questions, in fact this may be of no particular consequence. If it turned out that some of the questions were a bit harder, then it just means that the overall test score was a bit lower, but each student's score *relative* to the 59 other students would be the same. On the other hand, if we were administering the test on different days and wanted to make sure the students did not pass the answers along, we might want to get a new set of questions for the second day. If we did that, the extent to

which the overall difficulty of the test changed from day-to-day (the main effect of item) would be a source of error and would have to be considered in the reliability. If it did, we would be interpreting each student's score in an *absolute* sense. These ideas of relative and absolute errors arise again when we get to G theory, but are fairly easy to comprehend in this simple example.

This approach to reliability is commonly called "CTT" and has roots back to a famous book by Fisher (1925), called "Statistical Methods for Research Workers". Chapter 8 is titled "The Intra-class Correlation Coefficient". However, the basic ideas predate the book by a quarter of a century, back to Pearson (1896). The fundamental idea, leading up to the reliability coefficient, is that every score contains two elements, a true score and an error score. From this it follows that the reliability is just true score variance over total score variance, as we showed above.

What is wrong with classical test theory?

Let us go back to our example of the competence test. The performance on each task is determined by a number of factors:

Candidate factors

- knowledge;
- technical skills;
- social skills;
- intelligence; and
- charisma.

Task factors

- required knowledge;
- required technical skills;
- required social skills;
- required problem solving skills; and
- personality of standardized patient.

Rating factors

- rater expertise;
- rater severity.

Situational factors

- time of day;
- environment.

True random factor

We could go on and on. In short, there is not just one error affecting a score; there is a whole bunch of possible errors.

The trouble with CTT is that it is based on this very simple idea that any measurement has just two parts – true variance and error variance. But as we have just shown, there are multiple sources of error – some may be big and some may be small. But in CTT, we can only deal with them one at a time. So if we want to look at the effect of raters, we get different raters to score a bunch of tests and we compute inter-rater reliability. Or we could get the same rater to do it again, so we have two times, and calculate intra-rater reliability (note – and a different one for each rater, since we cannot combine raters and times). We might have a number of items on the test, and so we would calculate an internal consistency (reliability across items). And on it goes. The trouble is that we end up with a bunch of numbers between 0 and 1, but that is not really what we need. What we want to know is what are the big sources of error, so that we can get a lot of samples of them, and what are the little sources that we do not have to worry about. But since every study will typically use a different sample, with different raters and so on, we have no easy way to compare one to another.

What we really want to do is put all the sources of variance into one big analysis so that we can see what is introducing a lot of error and what is not, and run one big ANOVA that systematically computes all the error variances. We can then, if we are cunningly clever, determine different intra-class coefficients corresponding to inter-rater, test–retest, internal consistency, etc. And that is what G theory does (Cronbach et al. 1963).

However, in order to break the shackles of CTT, G theory begins with a new terminology reflecting, at its core, a new way to view the world of measurement. So here we go:

Basic concepts

The object of measurement (facet of differentiation). Most measurement situations are like our example, where we are attempting to see how well a particular instrument can differentiate between people – students, patients, or teachers. So our "object of measurement" – the thing we will ultimately attach a number to – is a person. As a result, most textbooks (Shavelson & Webb 1991; Brennan 2001) refer to any variance associated with the object of measurement as "person" variance.

However, some caution is necessary. If we have people rating hamburgers or wines, patients rating the food on the ward, or students rating textbooks, the object of measurement is actually the hamburger, the ward, or the book. So in these situations, "persons" can be books or hamburgers.[2] Also, the same design may yield different "objects of measurement" (Allal & Cardinet 1976). So, for example, we may be getting a group of patients to rate the seriousness of various health states. If we are interested in obtaining a number for each state, then the object of measurement is the written health state and the patient is a rater of the state. On the other hand, perhaps we are interested in the individual's optimism or pessimism about health in general. In that case, the patient is the object of measurement and the written health state is an item.

To remind researchers that the object of measurement is not fixed, and not necessarily human, but rather is the thing we want to ultimately be able to distinguish, Streiner and Norman (2008b) used a different terminology. They call this the "facet of differentiation.

Facets of generalization. G theory was developed by Cronbach et al. in a book published in 1972. A starting point was the recognition that the "true" score, derived from a hypothetical population (as we described in "signal and noise" section) was never observable and could only be approximated as the average across all the observations. So instead of clinging to this concept, G theory begins by defining a finite

"universe" consisting of observations across all the possible levels of all the factors the researcher is interested in. For example, if we were interested in estimating the contribution of raters, occasions, and cases to a measurement of communication skills, we define our universe in terms of a number of levels of rater, case, and occasion. So the "universe" score is the average score of an individual across all levels of all the factors in this specified finite universe.

We have been calling them "factors", but the accepted term in G theory is "facet". That is, in our study above, there are three obvious facets – rater, case, and occasion. However, although the terminology changes and the approach are much more versatile than CTT, it does retain some of the common features. In particular, we will still be calculating various ICCs, which are ratios of variances. And they have the same form as the traditional ICC, with the numerator representing the variance of interest (the "true" variance) and the denominator representing the sum of variance of interest and error variance. Not surprisingly, if we do a G study with only one facet – say, raters – the resulting G coefficient is just what would come out of the classical inter-rater reliability analysis described in "two-way ANOVA" section. And the interpretation is the same – a number between 0 and 1 representing the proportion of the total variance in the observations due to differences between the things we are trying to measure – the facet of interest. Like the classical coefficient, it is an index of the ability of the instrument to discriminate or differentiate between the things we are trying to measure

As we have already seen, in contrast to CTT, G theory allows for multiple sources of variance, which Cronbach calls the "*facets of generalization*". These are the facets that we want to generalize over (obviously). But this too requires a change in language:

- instead of saying, "What is the inter-rater reliability of this exam?" we say, "To what extent can we generalize these exam scores across raters?";
- and test–retest reliability becomes "To what extent can we generalize these scores across occasions?";
- and then a new one pops up, "To what extent can we generalize these scores across both occasions and raters?" which is sort of a test–retest inter-rater reliability.

But here is where G theory acquires its additional power. We are no longer constrained to the traditional variables like rater or time. Instead every measurement situation should be examined *de novo* by beginning with the question, "*What are the most likely sources of error in this particular measurement situation?*" If we do this, we are led into considering all sorts of variables (facets) that are not part of the usual lexicon, like cases and formats. Second, once these have been defined, we have essentially defined our finite "universe" of observations, and can determine the contribution of each facet to error, as well as the overall ability to differentiate objects across all levels of the universe.

How do we distinguish between those facets we want to generalize over (raters, in the case of inter-rater reliability) and those that we wish to hold constant (occasions, in this example)? This awaits the calculations, but for the moment, it is important to note that these have different labels. The facet we wish to generalize over is called a "random facet of generalization", signifying that we want to look at generalizing to some random, other level; the other facet(s) are called "fixed facets of generalization" – we hold them fixed. We will see in due course how these are dealt with in detail, but the basic idea is that fixed facets contribute to the variance of interest (true variance) and random facets contribute to the error variance. Logically, it is as if the fixed facets replicate the conditions of the original study and the random facets are a sample of a "universe" of possible allowed conditions.

Streiner and Norman (2008b) regrettably use a different terminology. In fact, they use the same terminology for different purposes. In their lexicon, a "fixed facet of generalization," as defined above, is a "fixed facet", and a "random facet of generalization" as above is just a "facet of generalization." That is, for an inter-rater reliability study, you would keep "item" as a fixed facet, and rater would be a facet of generalization. To make matters worse, they reuse the concept of fixed and random in another context, as discussed in the "Research Design" section.

Stratification facets. Now if you look in the textbooks on G theory, you will see that this is all there is. However, as we developed G_ String, we found that a very common situation in medical education is not adequately represented by this three – level world view. Imagine, for the moment, that the 60 students in our multiple choice test actually come from two classes, with different teachers. Students 1–35 are taught by Dr A; 36–60 by Dr B. As it turns out A is a much better teacher, so on average the first 35 students do a lot better than the other 25. We really have another facet in the design, call it "classroom".

What kind of facet is it? Not a facet of generalization – clearly we are not trying to generalize from a student's performance in one class to her performance in the other one, since she can only be in one class. But somehow it matters, since the first 35 students are getting a score which is biased upwards compared to the last 25.

It really is a different kind of facet altogether, which we can identify by analogy. In designing experiments, we often sample students in different *strata* – gender, educational level, classroom, and school. We are doing the same thing here, although the intended use differs, but consistent with the analogy, we will describe these as "*stratification facets*". Again, the specific approach to dealing with stratification variance will be dealt with later.

Summary

In this section, we have reviewed some basic concepts about CTT, pointed out its weaknesses, and begun to show how G theory deals with these problems. The essential difference is that, instead of simply dividing an observed score into true score and error, G theory explicitly identifies multiple sources of error (facets). Facets are of two kinds – the facets of generalization (errors) and stratification facets (strata or groups).

Research design

Outline

- G study research designs
- G coefficients – General form
- Constructing G coefficients
  - Absolute *vs.* relative error
- D studies – G coefficients for multiple levels

G study research designs

We have talked a lot about G theory and its relation to variance components – at least at a conceptual level. It is now time to get a bit more practical. In "two-way ANOVA" section, we showed how to compute true and error variances for a simple CTT design; what we might call in G theory a "one facet" design, referring to the single facet of generalization. As you recall, we used repeated measures ANOVA, with a single repeated "within-subject" factor. In G theory, we take the same basic approach, but extend it by introducing additional repeated measures in the design.

Let us go back to the assessment of communication skills. We will do it as many people do, by creating an Objective Structured Clinical Examination (OSCE), where the student goes from room to room interviewing standardized patients, moving on every 10 min or so. And when she leaves the room that standardized patient completes some kind of communication skills rating. Let us say we have 10 stations. If that is all there was, we could treat the whole thing as a simple reliability study and use CTT, as we did in "two-way ANOVA" section. We would have 10 repeated observations on each student, and we would set it up as a one-factor repeated measure ANOVA.

But let us recast the whole thing as a G study. To begin with, the *object of measurement* is straightforward – student. That is, after all is said and done, we are interested in seeing how well the test differentiates among students – consistently identifies high and low scoring students. Following our earlier lead, let us assume that we have three sources of error – *facets* – of interest: raters, cases, and occasions.[3] We will include all of these in a single design. We have the 10 cases (stations) in the original design. We also want to see how much error is coming from different raters, so we would perhaps include a second rater (another standardized patient perhaps) as an observer in each station. We now have a two-factor repeated measures design, with rater (2 levels) and station (10 levels) as the two facets of generalization. Finally, perhaps student skills vary from day-to-day. To test this, we might actually create a 2-day test, so they do five cases on one day and then again two weeks later. Now, we have a three-factor ANOVA with day, case, and rater as facets of generalization.

There is a subtle difference between these facets that has a major impact on the design and analysis. In ANOVA terminology, day, station, and rater are all said to be *crossed* with student; by that we mean that every level of day or rater occurs at all levels of student – each student does both days, all five cases, and all raters. In the design above, station and day are *crossed* as well, since each case occurs on both days. However, rater is *nested* in station, since each rater occurs at only one station. We could have made station nested in day by using 10 different stations, 5 on day 1 and 5 on day 2. And we could have, with difficulty, crossed rater with station by, for example, video-taping each performance and having the same two raters watch all students and all stations.

We can also introduce stratification facets into the design. Commonly, OSCEs are done with multiple circuits so a large number of students can be "processed". If we did that, then we could imagine that the first 10 students do circuit 1, 11–20 do circuit 2, and so on. In that case, assuming all circuits use the same stations, then rater and student are nested in circuit. And since student is the facet of differentiation, circuit becomes a stratification facet.

Why does all this matter? As we increase the number of factors/facets, we not only increase the number of main or overall effects in the design, but we also introduce various interactions. For example, a station × day interaction is a measure of the extent to which the scores assigned to different stations are different on successive days. A student × station interaction shows whether different students find different stations easy or hard. The magnitude of various interactions provides useful information about sources of errors. In addition, all of these interactions have to be dealt with individually in constructing G coefficients, as we will begin to see in the following section.

However, we can only have an interaction between crossed facets. As a counter-example, while we can have a student × day interaction (do some students do better on day 1 and some on day 2?) we cannot have a student × circuit interaction (do some students do better on one circuit than another?) since each student can only be in one circuit. And we cannot have a rater × station interaction for a similar reason. Nested facets are written as student : circuit and rater : station to make this distinction.

Finally, we can use this example to illustrate another bit of technical jargon. So far, the design is *balanced* – there are 10 students per circuit for all circuits, 5 cases, 2 raters, etc. But suppose some students were overcome with nerves and dropped out, so that the number of students per circuit varied from 7 to 10. We would now say the design is *unbalanced*. While G theory software is capable of dealing with unbalanced designs of considerable, though not unlimited complexity, the multipliers for individual terms in the G coefficients becomes more complex. We will get into the details later.

G coefficients – General form

To this point, we have discussed various kinds of facets – the object of measurement, which is roughly equivalent to "subjects" in CTT, facets of generalization, which are equivalent to the various other facets such as "raters," "items," or "occasions" in CTT, and "stratification facets" that have no classical equivalent. We now describe general strategies to combine these into G coefficients.

Every G coefficient has the same form of an intra-class correlation as described in "signal and noise" and "two-way ANOVA" sections. As before, it is a ratio of the "true" variance to the sum of "true+error" variance. What differs is the individual variance components that go into the numerator and the denominator.

We have already distinguished between the variance due to the object of measurement (and there is only one of them) and the facets of generalization. For consistency with Brennan, we will always describe the facet of differentiation as *p* for "persons." But as in the OSCE example, we can have any number of facets of generalization limited only by imagination and logistics. In turn, these can lead to distinct G coefficients, so that a single G study design can theoretically create a large number of coefficients. Some may be analogous to classical coefficients, such as inter-rater, test–retest, or internal consistency; others may have no classical equivalent. For a crossed design with 2 facets, we can have 3 G coefficients; for 3 facets, 7 coefficients, for 4 facets, 14 coefficients, etc. Of course not all these are interpretable or useful. For obvious reasons, we cannot come up with standard names like "inter-rater" to describe these various coefficients. Instead, we adopt a standard nomenclature, in which we distinguish between *fixed* facets of generalization and *random* facets of generalization.

To see how this works, let us return to the OSCE again. We have P(erson) as the facet of differentiation, S(tation), D(ay), and R(ater) as facets of generalization. Below is the complete set of possible G coefficients, with classical equivalents where they exist:

	Facets of generalization			
Object of measure	Fixed	Random	Question	Classical equivalent
P	D,S	R	To what extent can I generalize from one rater to another?	Inter-rater
P	D,R	S	To what extent can I generalize from one station to another?	Internal consistency
P	R,S	D	To what extent can I generalize from one day to another?	Test–retest
P	D	R,S	To what extent can I generalize from one rater on one station to another rater/ station?	
P	S	D,R	To what extent can I generalize from one rater and day to another?	
P	R	D,S	To what extent can I generalize from the same rater on one day and station to another?	
P		D,S,R	To what extent can I generalize across all facets to a comparable overall Test?	

While the generalizations over two facets do not make much sense, the last coefficient certainly does. It basically says, "*Given the way we have defined the universe of observations, considering all the sources of error, this is the coefficient indicating how well we can generalize any score to another parallel test*." Further, the other coefficients can be directly compared to determine large *vs.* small source of error, since all are based on the same sample of subjects, cases, raters, etc. In short, we can use the information from the G study diagnostically to identify major and minor sources of error.

How are the coefficients constructed? We construct a *signal* term and a *noise* term, very much like the preceding discussion. The signal, labeled τ (tau) by Brennan, consists of all the variance components (main effects and interactions) due to the object of measurement and all the fixed facets. The noise term labeled either Δ (DELTA) or δ (delta) comprises all main effects (Δ) and interactions (Δ and δ) that have the facet(s) of generalization in them (we will go into this in more detail later).

The distinction between Δ and δ relates to the absolute or relative error coefficients. If we wish to interpret a person's score relative to all other persons in the study, the main effects of the facets of generalization (e.g., the main effect of "item") are irrelevant; it amounts to moving everyone up or down by the same amount. So we would omit any main effects of G facets from the error term. Conversely, if we want to place an absolute interpretation on a person's score (John's IQ is 122), then any systematic (main) effects of item or rater introduce error into this estimate. These main effects (and the interactions between G facets) now go into the error term.

If we are computing the absolute error, we use Δ; if we are calculating relative error we use δ. Basically, Δ contains all the relevant main effects; δ does not. So, conceptually, for each G set of facets of generalization and differentiation, we have two coefficients:

$$E\rho^2 = \tau/(\tau+\delta) \quad \text{relative error,}$$
$$\Phi = \tau/(\tau+\Delta) \quad \text{absolute error.}$$

Unfortunately, Streiner and Norman (2008b) use a different approach. They approach the issue of absolute *vs.* relative error one facet at a time, and each facet is identified by the user as either absolute or relative as part of the data input. For example, if we were analyzing a clinical skills rating form, to be completed by different observers, we might well decide that we can ignore the main effect of item, since all students are rated on the same items. The main effect then just shifts every person's score up or down by the same amount. In Brennan's terminology, we would say that we would use relative error, since this amounts to excluding the main effect of item from the denominator. On the other hand, if each student is rated by a different supervisor, then any main effect of rater would affect different students differently, so should be included – absolute error. Brennan does not allow this individualization, although in "computing G coefficients" section, we will discuss how to create coefficients reflecting these differences. Streiner and Norman (2008b) do allow it in their description, however they use different terminology. If the main effect should be included, this amounts to declaring a "*random factor*," if not, the effect is declared a *fixed facet*. The terminology originates in the idea of fixed or random effects in ANOVA.

Some other examples of absolute and relative errors. In a self-completion scale (learning style or depression) we would likely use relative error, since the items are always the same. If we had an essay test where people could, for example, write on three topics of the five listed, we would use absolute error, since the main effect of essay is reflected in differences on individual scores.

D studies – G coefficients for multiple levels

In CTT, one useful extension is something called the Spearman–Brown formula which is used to determine the reliability of a test that is "*k*" times as long as the original test. That is, if the study computed the reliability of a 10-item test, we could use the Spearman–Brown to estimate how reliable a 20, 40, or 100-item test would be. The basic strategy amounts to dividing the error variance term by "*k*" (analogous to the relation between the standard deviation and the standard error of the mean).

G theory goes two better. Because we are simultaneously considering error variance from multiple sources, we can also vary the number of levels of each source. In this example, having determined the generalizability of a single observation over raters, stations, and days, we can first determine the generalizability of the test we used by dividing error variances containing day by 2; containing rater by 2, and containing station by 5.

But the next extension is even more interesting. Once we have the variance components, we need not stick to the levels in the original study. We can insert whatever "*n*s" we choose. In this *decision* or *D* study, we can then ask optimization questions like, "Given that we have $2 \times 2 \times 5 = 20$ observations available, what combination of observations yields the maximum generalizability?" We could investigate a number of possibilities:

Rater	Station	Occasion
2	5	2
4	5	1
1	10	2
2	10	1
5	2	2
2	2	5

When you think about it, the general strategy would be to spread the observations out so that facets associated with large error variance are divided by larger numbers and those with small error variance are divided by small numbers. But you do reach a law of diminishing returns, so that typically the optimal generalizability occurs with intermediate values.

Summary

In this section, we have introduced the reader to the basic concepts of G theory. We have shown that it is an extension of CTT that deals with multiple sources of error simultaneously. We have illustrated how this permits much greater precision in examining and reducing the sources of error variance in a measurement situation.

Designing your G study

Outline

- What is the dependent variable?
- What is the "object of measurement?"
- What are the facets of generalization?
- What (if any) are the stratification facets?
- Which facets are nested and which are crossed?
- How do I specify the number of levels of each facet?

Designing G studies is a bit tricky, but it basically follows the same steps you would use in any research study. You must figure out your dependent variable, independent variables, and research design. There are constraints on the choice and configuration of variables, which we will get to in due course. But the steps remain the same.

What is the dependent variable?

What are you trying to measure? Is it a score on a written test? A rating scale filled out by a supervisor? Are you interested in biological measurements like body mass index or range of motion? G theory is pretty well indifferent to the actual measurement. In particular, this is the right time to deal with some longstanding myths about measurement.

Myth 1: Rating scales are "ordinal measurements" and you have to use non-parametric (like Spearman's rho or chi-square) to analyze them.
If this were so, G theory would be out of business, since it is based entirely on ANOVA which is a "parametric" procedure. This particular axiom has been around as long as we have been doing measurement (we, the authors) which is a pretty long time. And it has been disproved again and again for almost as long (Norman 2010).

Myth 2: Your data have to be normally distributed or you cannot do ANOVA.
You do NOT have to have normally distributed data. Anyone who says so is revealing his ignorance of basic statistical theory. ANOVA and similar methods are based on distributions of *means*, and the central limit theorem says that, for moderately large sample sizes (>10), the means will be normally distributed regardless of the actual distribution of the data.

So you can do G theory analysis on just about any kind of measurement. In fact, you can even do G theory on sets of 0's and 1's (like dead = 0, alive = 1). This sounds bizarre in the extreme; you are supposed to use Cohen's Kappa (Cohen 1960). However, in 1973, Fleiss and Cohen showed that Kappa and the ICC (which we described in "two-way ANOVA", and underlies all of G theory) are mathematically identical. On more than one occasion, we have done G analysis on binary data, computed G coefficients, and reported them as "generalized kappa's".

It sounds like there are very few constraints, and there are. But there is one that really matters. As we have already described, G studies depend on repeated measurements across different conditions (e.g., three raters, two times), to partial out the error variance. And all the variance derives from deviations from the overall or "universe" mean. For this to happen, it must be meaningful to average across all conditions. So multiple raters using the same scale is fine. But if we were looking at, for example, the student's total score in a course, where the individual subscores were, (a) 2 assignments based on 10-point scales, (b) group participation out of 5, (c) a midterm out of 20, and (d) a final exam out of 60; it would be nonsensical to average these. If you converted everything to a percent, let is say, then statistically it may be defensible. But even so, conceptually, does it make sense to talk about

the generalization from an assignment score to group participation? Perhaps; but all may be samples of an underlying trait called "statistical competence" – the issue requires careful thought.

What is the ''object of measurement''?

This seems about the easiest question of all. What are we trying to attach the measurement to? In the above example, it is student; we are trying to figure out how well our various tests can differentiate among students (tell them apart). But perhaps because it seems so straightforward, every so often someone gets it entirely wrong. For example:

(a) We want to get patients to rate their satisfaction with their experience on our inpatient ward. The object of measurement is "ward". Patients are raters of the ward. And the study, as conceived, is undoable, since with only one ward you clearly cannot work out the variance between wards.
(b) We get our friends ($n=10$) to rate a number of red wines ($n=5$) on four seven-point scales.

There is always one and only one object of measurement per analysis. Once this is decided, you can now take the first step in creating the study design and database. Imagine the spreadsheet that will contain all the data. The start is to assign one line (row) to each subject (object of measurement), so that all observations for each subject will then fill in successive columns. Subject ID can be created for each row; the exact numbering approach is irrelevant as it is not needed by the analysis software. Most G studies will follow this "one line per subject" strategy, but not all. We will identify the exceptions later.

Because the object of measurement is usually a person, most books on G theory reserve the symbol "p" for the object of measurement. We will do the same when we get to computing G coefficients.

What are the facets of generalization?

Well, now is the time to let your imagination run wild. What are the possible sources of error in any estimation? Some obvious ones come easily to mind – raters (of communication skills, say) hence inter-rater reliability; items (on a test) – internal consistency; times or occasions, and test–retest reliability. But if we have a test comprised of multiple cases, a much more important source may be the cases. Inter-rater reliability is usually around 0.7–0.8, however study after study shows that inter-case reliability is closer to 0.1–0.3. That is why OSCEs work; they assess competence by averaging over anywhere from 10 to 20 cases or stations.

Other facets are less obvious. Often rating scales have multiple subscales – quality of life may be comprised of physical, social, and emotional functions. These dimensions can be viewed as a facet of generalization as well. Different experimental conditions may be a facet; for example, to what extent can we generalize a rating of a surgical resident's suturing skills from a static simulation to an actual patient.

There is an obvious upper limit in that we can only make a limited number of assessments before our subjects' rebel. So you should try to arrive at a list that encompasses all the likely large error sources. Once we have identified these facets, we are well on the way to designing the study. But first we must examine one other kind of facet.

What (if any) are the stratification facets?

In the Research Design section, we discussed the idea of stratification facets. These arise when the facet of differentiation (usually persons) is "nested" or "blocked" in some other variable. Examples abound:

(a) Tests may be conducted at different locations or on different days, so student is nested in location or day.
(b) Patients may have different severity of disease, so if we are seeking a measure of anxiety or depression, we may want to stratify patients by severity.
(c) We may be conducting a validity study where we will examine whether senior students perform, on average, better than junior students; student is stratified in educational level.

Returning to the database, we can envision the stratification facets as additional columns. These can be identified by some ordinal index, most commonly 1, 2, 3,.... *G_String requires that all subjects within a particular stratum occur together*. So the first $n1$ rows may be from the first stratum, the next $n2$ rows, from the second stratum, and so forth.

Which facets are nested and which are crossed?

We defined "nested" and "crossed" in "research design" section, but let us review. A facet "A" is crossed with "B" if each level of A occurs at all levels of B; it is nested if each level of A occurs only at one level of B. So what, exactly does that mean? It comes clear with a few examples.

(a) An inter- and intra-rater reliability study in radiology. Three raters examine 50 chest films on one occasion and again two weeks later. There are two facets – rater and occasion. Since each rater is present on both the first and second occasion, rater is crossed with occasion.
(b) Forty students go through an OSCE with 10 stations, each with 2 raters. Unlike most OSCEs, all stations use the same four-item rating form with Likert scales. There are three facets – station, rater, and item. Rater is nested in station, since the same raters cannot be present at all stations (they could, we suppose, if raters and students went along together from station to station. But this design would completely confound rater differences with student differences and would be a very bad idea). Item is crossed with station, since all stations use the same items. If it were the more common checklist specific to each station, then item and rater are both nested in station.
(c) Seven teachers of different sections of a first-year psychology course are rated by their students on the

same rating scale. Teacher is the facet of differentiation, and here student (rater), a facet of generalization, is nested in teacher since each student belongs to only one section. But item (on the scale) is crossed with teacher, since all teachers are rated on the same items.

(d) Fifty patients complete a quality of life scale twice 2 weeks apart. It contains three subscales: physical function with 12 items, social function with 8 items, and emotional function with 6 items. There are three facets of generalization – time, (2 levels), subscale (3 levels) which are crossed (all subscales occur on both times), and item, which is crossed with time but nested in subscale, and has 12, 8, and 6 levels.

From these examples, it is evident that we can have various kinds of nesting – one facet of generalization nested in another facet of generalization, (b and d) or a facet of generalization nested in a facet of differentiation (c). Further, nested facets may have the same number of levels in each nest (two raters per station in B) or different (raters in c; items in d). If it is the same number, it is called a "balanced" design; if not, an "unbalanced" design.

One thing we cannot do, a limitation of urGENOVA, is have a facet of generalization nested in a stratification facet (such as students at different hospitals having different OSCE stations).

Specifying the number of levels

How do we deal with unbalanced nested designs?. Setting up a crossed design is easy. In the wine-tasting study we have 5 wines × 10 raters × 4 items. We would likely just create a single row for each wine and spread the $10 \times 4 = 40$ ratings across 40 columns.

Balanced nested designs are also easy. In example (a) above, we have a total of four facets and all but rater are crossed: student (40 levels), station (10 levels), rater (2 levels in each of 10 stations), and item (4 levels). Again, we would likely just layout the data in one record per student, with $10 \times 2 \times 4 = 80$ columns. Alternatively, we could have one line (row) per station, so that the first line is student 1, station 1; line 2 is student 1, station 2,.... Line 11 is student 2, station 1.

In nested designs, when G_String gets to asking for the levels, it will create as many boxes as there are nests (in this case, 10). When you input the first value (2) it will automatically create the same number in all the remaining cells. If the design is balanced, then all levels are now specified. However, if the design is unbalanced, (i.e., the number of raters per station varies) then you can overwrite these automatic values.

There is another way, however. Situations arise where there are large numbers of "nests". In one study, there were 15,000 patient ratings of 1000 physicians; the number per physician varied from 2 to 35. In another, students rated lectures from different faculty members; the total number of ratings exceeded 27,000 and varied from 1 to 55. These are both situations where rater (patient or student), a facet of generalization, is nested in teacher or doctor, the facet of differentiation. If we were to do the problem manually, we would have to enter over 1000 two digit numbers into G_String, with real possibility of error. The alternative is that G_String will do this automatically. We simply create an index for teacher or doctor (an index for rater is not necessary) and ensure that there is one record *per rater* and all record are sorted in ascending order, so all ratings of doctor 1 occur first, then doctor 2, then doctor 3, and so forth, G_String then automatically generates counts of the numbers in each nest.

We can use the same automatic facility if we have a nested, unbalanced facet with multiple observations on each record. For example, we get people leaving hamburger shops to rate each burger on a single seven-point scale, and we get from two to seven ratings. The facet of differentiation is "Burger;" and there is one facet of generalization, rater. If these two to seven ratings are all on the same row, G_String can automatically count how many ratings there are for each burger.

Summary

If you follow these steps, you will now have all the information you need to design the G study and to devise the format of the database.

Computing G coefficients

Outline

- G coefficients
- General rules
- Rules for creating τ, δ, and Δ
- Rules for creating the divisor for each facet
- Absolute, relative, and mixed error coefficients
- Coefficients for nested designs

In the last section, we conceptually worked out the design of the study. It is now ready to be analyzed; the specifics of how G_String is set up for analysis are described in the following two sections. Strictly speaking and from a purely practical point, you do not actually need to read the following explanations. However, if you are suffering from an irrepressible need to understand what you are doing, here it is. In this section, we will go through the theory of calculating G coefficients. The seven steps of generalizability analysis are:

(1) formalizing the problem;
(2) organizing the data;
(3) calculating group means;
(4) calculating mean-square differences for groups;
(5) estimating group variances;
(6) estimating variance components for effects; and
(7) calculating appropriate ICCs.

Steps 1–6 represent standard analysis of variance approaches. Steps 3–6 are actually being calculated by urGENOVA within G_String with no fuss, so we need not go into much further detail. The specifics are all handled in standard statistics texts that deal with repeated measures

ANOVA. The preceding sections gave you an overview. However, to go from variance components to G coefficients in Step 7, we are treading onto unfamiliar territory. Most of the details have been worked out by Brennan, and what follows is a direct interpretation of his theories. We have added some extra theory around stratification facets, which we will describe in due course.

To begin with, let us remind you of the general form of the G coefficient:

$$G = \frac{\text{var(signal)}}{\text{var(signal)} + \text{var(error)}}.$$

What happens in G theory is that, although the total variance (var(signal) + var(error)) remains the same, as we make some facets fixed and other facets random, we actually change the apportioning of this variance to signal and error. To make the algebra a bit more efficient let us redefine the terms, consistent with Brennan. Signal is called τ, and error is called δ if it is relative or Δ if it is absolute (See "research design" section). So the general form of the G coefficient, in Brennan terminology, is:

$$G = \frac{\text{var}(\tau)}{\text{var}(\tau) + \text{var}(\Delta \text{ or } \delta)}.$$

We now have to work out what specific variance components (main effects and interactions) go into each of τ, Δ, and δ. In the rules below we develop a framework that applies to any design. It is based in part on Brennan's (2001) rules on p.122 of his book.

Rules for assigning types of facets

Rule 0[4]

Components of variance result from three sources:

(a) the object of measurement (facet of differentiation), p. There is only one p (for 'person'),
(b) facets of stratification,[5] S_1, S_2, ...These are of the form $p: S_1,S_2$, defined in Screen 4 of G_String, and
(c) facets of generalization: G_1, G_2, G_3, ...

Rule 1

Facets of stratification (S_i) appear in ANOVA (and eventually in G_String), but cannot be facets of generalization. Facets of stratification can be recognized by the fact that they provide containers for a nested object of measurement (facet of differentiation).

Rule 2

Facets of generalization, either nested in or crossed with p, are specified as of two types in the calculation of G coefficients: random facets R_j, and fixed facets F_k. These are specified in Screen 12 (and can be changed by the user on successive calculations). In the intitial run run, G_String automatically sets all facets of generalisation to random.

Rule 3

Nesting of variables may arise in several different ways and are handled differently according to the rules to follow.

(a) $p{:}S_i$ – by definition, p can only nest in S_i. These are handled in Rule 1.

> For example, when programs run an OSCE, it is very common for students to be nested in hospital, day, or circuit. National examinations may have candidates nested in city.

(b) $G_i{:}p$ – facets of generalization can be nested in the object of measurement.

> For example, students' ratings of teachers, or patients' rating of doctors, where each student or patient has only one teacher or doctor, but each doctor or teacher may have multiple ratings by different patients or students.

(c) $G_i{:}G_j$ and $G_i{:}G_j{:}G_k$ – facets of generalization can be nested in other facets of generalization, such as items nested in OSCE station.

> One example is an OSCE using checklists, where each station has a different content-specific checklist. Another is a case-based written examination, where each case may have several questions, so question is nested in case. Questionnaires with subscales often have item nested in subscale.
> Another example would be items within subscales (e.g., verbal reasoning, analogies) in an IQ test. Or individual questions nested within cases in a written patient management test.

Note: *Nesting of facets results in elimination of certain interactions in the ANOVA, but these are handled automatically by urGENOVA. There are also implications for the division by "n" in D studies.*

Rules for creating the variances of τ, δ, and Δ

Rule 4 (Brennan, 2001, Rule 4.3.1, p. 122).

$$\begin{aligned}\text{var}(\tau) = {} & \text{var}(p, \text{including } p : S) \\ & + \text{var}(\text{all } p \times F_k \text{ interactions not containing any } R_j) \\ & + \text{var}(\text{all main effects of form } F_i : p \text{ not containing any } R_j).\end{aligned}$$

Explanation: The general strategy is that τ contains the object of measurement and all its interactions with fixed facets. The reason behind this rule with respect to nested variables is that with fully crossed design, τ contains all interactions between p and F but not the main effect of F. With nested design, the variance due to nesting (e.g., $\text{var}(F_i{:}p)$) actually contains the $p \times F_i$ interaction so is in τ term.

When a facet is a facet of generalization, its main effect will be in Δ. However, when it is a fixed facet, (if it is not nested in p as below), the main effect does not move to τ, see Brennan (2001, section 4.4.1). *He states that "fixing a facet affects which variance components contribute to τ and δ but it does not change their sum." However, in the example it DOES change sum of τ and Δ since, when facet is random its main effect is in Δ but when it is fixed, main effect is not in τ.*

All effects that contain the facet of differentiation but no random facet of generalization contribute to $\sigma(\tau)$.

Rule 5 (Brennan, 2001, Rule 4.3.3, p. 122).

$$\text{var}(\delta) = \text{var}(\text{all terms containing } p \text{ and } R_j\},$$
$$\text{including specifically all terms of form}$$
$$p \times R_i \times R_j, R_j : p, p \times R_j : F_i, p \times F_i : R_j, p \times F_i \times R_j).$$

Explanation: The error term consists of all terms containing the random facet(s), R. The reason behind this rule with respect to nested variables is that, with fully crossed design, δ *contains all interactions between p and* R_j. *With nested design, the variance due to nesting (e.g.,* $\text{var}(R_j{:}p)$, $\text{var}(p \times F_i{:}R_j)$*) actually contains the* $p \times R_j$ *interaction* $(R_j + R_j \times p)$ *in the first case,* $p \times F_i + p \times F_i \times R_j$ *in the second case and, therefore, belong into the error term.*

All effects that contain at least one random facet of generalization and interactions between the random facet(s) and the object of measurement, p, contribute to $\text{var}(\delta)$.

Rule 6 (Brennan 2001, Rule 4.3.2, p. 122).

$$\text{var}(\Delta) = \text{var}(\text{all terms containing } R_j)$$
$$+ \text{var}(\text{all terms containing } S_j,$$
$$\text{specifically including all main}$$
$$\text{effects of } R_j, \text{ all interactions of form } p \times R_j)$$
$$+ \text{var}(\text{all interactions between } R_j \text{ and other}$$
$$\text{facets, e.g., } R_jF_k \text{ and } R_jR_k)$$
$$+ \text{var}(\text{all terms containing } S_j \text{ to left of}$$
$$\text{colon including main effect of } S)$$
$$+ \text{var}(\text{all interactions between } S_j, \text{ and}$$
$$G \text{ facets; but excluding terms where}$$
$$S_j \text{ is to the right of the colon}).$$

Explanation: All effects that contain at least one random facet of generalization and all effects that contain a stratification facet (unless the S facet is to the right of the colon) contribute to var(Δ).

Stratification facets are of two types – those that might be termed "experimental," where there is an anticipation that there will be a large main effect of the stratification facet (e.g., educational level) and those that are part of the design, but the expectation is that there would be no effect of S (e.g., day, circuit, in an OSCE). For experimental strata: (a) the remaining facets are crossed (every stratum gets the same measures (for example, all persons get the same test items)) since this is the only way that one can test hypotheses about differences, (b) the RELATIVE ERROR term is appropriate, as generalization is within facet. For design facets, the strata may contribute error in interpretation of particular scores, so the appropriate term is the ABSOLUTE ERROR term.

Rules for creating the divisor of each facet in the G coefficient

Rule 7

For balanced designs, the divisor of each facet in a term in τ, δ, or Δ (except for terms involving p or S) is the number of levels of the facet in the term. For terms in p or S the divisor is always 1. For every interaction term in τ, δ, or Δ, the divisor is the product of the divisors of the facets making up the interaction. Thus a term of the form G × H will be divided by the product of the divisors of the individual terms. So for crossed designs, terms like $g_1 \times g_2$ will be divided by $n_{g1} \times n_{g2.}$ For nested facets of the form $g_1 : g_2$ the divisor is also $n_{g1} \times n_{g2}$.

> For balanced designs, the divisor for each facet of generalizability is the number of levels of the facet. For p and S facets, the number of levels is always one. For nested facets of form $g_1{:}g_2$ the divisor is $n_{g1} \times n_{g2}$.

Unbalanced designs are of three types, and each requires different treatment.

Type 1: $p{:}S$ (Person nested in a stratification facet)

We have already dealt with this situation above. The divisor for a stratification facet is always one.

Type 2: $G{:}p$ (Facet of generalization nested in person)

In this design, at least one facet is nested within the object of measurement.

> A very common example is student ratings of teachers, where each student is in one classroom, so student is nested in teacher. "Multi-source feedback" is another; a physician seeks ratings from a number of patients and colleagues. Patients are nested in physician.

In this case, the G coefficient is divided by the number of levels of person (which is set at one) × the *average* number of levels of rater, which according to Brennan is computed as the "harmonic mean" – basically the average of the $(1/n)$ terms, one for each p.

$$\tilde{n}_g = \frac{n_p}{\sum_p \frac{1}{n_{g:p}}}.$$

For unbalanced designs of form $G{:}p$, the divisor is the harmonic mean of the number of observations at each level of G.

Type 3: $G_1 : G_2$ (One facet of generalization nested in another)

This situation, where one facet of generalization is nested in another, is encountered quite frequently. Some examples:

> – a case-based test, where each case has different questions (and different numbers of questions),
> – a questionnaire with different questions in each subscale, and
> – an OSCE, with different checklists with different numbers of items, in each station.

In this situation, as in balanced designs, there is an "n" associated with the appearance of $G_1 : G_2$ and another associated with G_2. For G_1, the "sample size" is simply the total number of observations of G_2, which Brennan calls "n_+"

(pp. 219, 232). For a balanced design, this is just $n_{G1:G2} \times n_{G2}$. For an unbalanced design, it is:

$$n_+ = \sum_{g2} n_{g1:g2}$$

For example, if 5 subscales had 2, 4, 5, 7, and 11 items each, $n_+ = 29$. If 5 subscales had 3 items each, n_+ would be $5 \times 3 = 15$.

However, the imbalance in G_1 also affects the denominator of G_2. According to Brennan (2001, p. 232), whenever G_2 appears, the variance will be divided by $\tilde{n}_{g2}$, which equals:

$$\tilde{n}_{g2} = \frac{n_+^2}{\sum_{g2} n_{g1:g2}^2}.$$

In the above example, this equals $29^2/(2^2+4^2+5^2+7^2+11^2) = 841/215 = 3.9$.

Type 4: $G:\underline{p}:S$, $G_1G_2:p:S$ (Facet of generalization nested in person nested in facet of stratification)

This design is an extension and combination of Type 1 and Type 2. Person is nested in one or more stratification facets and at least one facet is nested within the facet of differentiation.

As in Type 2, this may arise from a situation where students in a class rate their teacher, or employees in a company rate their supervisor. However, now the facet of differentiation is also nested (e.g., teacher within school or supervisor within company).

As before, the divisor for terms involving S is always 1, since each person can only occur at one level of S. Similarly, for the facet of differentiation, p, the G coefficient is divided by 1. And as before for the facets of generalization nested in the stratification and differentiation facets, the divisor is the harmonic mean.

Thus in the $G:p:S$ design, the specific terms are S, $p:S$, and $G:p:S$. Variance due to S and $G:p:S$ would be in Δ term; only $G:p:S$ would be in the δ term. The divisors for these terms in the G coefficient are $S/1$, $p:S/1$, and $G:p:S/\tilde{n}_g$ where, as before, $\tilde{n}_g$ is the "harmonic mean" – the average of the $(1/n)$ terms.

In the $G_1G_2:p:S$ design, the specific terms are S, $p:S$, $G_1:p:S$, $G_2:p:S$, and $G_1G_2:p:S$. As in the single G facet case, divisors for p and for S are 1; any term involving a facet of generalization (G) will be divided by the harmonic mean of the n's contributing to the variance. And the product term $G_1G_2:p:S$ is divided by the product of the two harmonic means.

$p:S\ g:S$ multiple stratified facets

There is one common design that cannot be handled in urGENOVA and, therefore, G_String. It is often the case that tests may occur at different sites or on different days. If different sites use different forms (or, more commonly, different raters) then, strictly speaking, the G facet is nested in the S facet. In our formalism, this is a $p:S\ g:S$ design. urGENOVA does not handle this case. Consider what would happen if we found a difference between sites. This could arise either because students at site A are better than those at site B, or raters are more lenient at site A than site B. There is no way of disentangling those two. Similarly, if there is no difference, the two sites could be equal, or any difference in student competency could be compensated by opposite differences in rater leniency These ideas are elaborated in Keller et al. (2010).

Absolute error, relative error, and mixed error coefficients

In "research design" section, we discussed the conceptual difference between the absolute error coefficient and the relative error coefficient, where the former included all main effects in the error term Δ, and the latter include no main effects in the error term δ. However, the situation may arise where it makes sense to include some main effects but not others. For example, if a group of medical charts are being audited by peer review using a standard form with 10 items, it may make sense to include the main effect of rater, since different charts may well be rated by different raters, so any rater bias may not be identifiable. On the other hand, since all raters are using the same checklist, the fact that some items may be systematically harder or easier than others is irrelevant to score interpretation.

This is not done automatically in G_String, but it is not difficult to use the variance components produced by G_String to create mixed coefficients. It also commonly emerges that the absolute and relative error coefficients are very similar, in which case further refinement may be unnecessary.

Summary

In this section, we have explained the rules for conducting G study analyses and for generating G coefficients for any (permitted) class of design.

G theory software

Outline

- From theory to software
- Overview of G_String

As we have seen previously, analysis using G theory involves a number of steps:

(1) formalizing the problem;
(2) organizing the data;
(3) calculating group means;
(4) calculating mean-square differences for groups;
(5) estimating group variances;
(6) estimating variance components for effects; and
(7) calculating appropriate ICCs.

There is no fundamental reason why any of these steps could not be performed by hand using nothing but paper, pencil – and endurance. In reality, however, homo iPhone either forgot how to do mental arithmetic, or she/he does not have the necessary patience any more. Some or other electronic gadget, therefore, has to come to the rescue.

Steps 1–4 are, apart from the numerical effort, relatively trivial. In the case of well-balanced datasets, Steps 5 and 6 do not pose conceptual hurdles either. What do we mean by well balanced?

- Facets and population are either mutually crossed, or where nested, the number of nesting levels in each facet stays constant throughout.
- There are no missing data.

Things get complicated, when either or both of these conditions are not met. As long as the imbalance is relatively minor, we can employ an analogous-ANOVA method. However, the random nature of actual datasets does not guarantee that all estimates for variance components come out positive. The convention for dealing with negative variance component estimates is to set them arbitrarily equal to zero, without adjusting the other component estimates accordingly. Other, more complex methods for estimating variance components with unbalanced datasets exist. But for the case of G studies, this does not make a significant difference.

Most general purpose statistical packages, such as SPSS, SAS, or MATLAB, can handle Steps 2–6, as long as the population size and number of facets are small enough. However, these programs attempt to solve general design matrices, whose dimensions grow very quickly with sample size and number of facet levels. As a consequence, calculations of generalizability parameters for realistic datasets can quickly become very time and resource consuming.

As a consequence, researchers have developed specialized computer programs for generalizability analysis since the 1980s. One of the first programs ETUDGEN, developed by François Duquesne at the University of Mons in Belgium around 1982, is the ancestor of EduG by Cardinet et al. (2010).

One of the pioneers in G theory, Brennan from the University of Iowa, developed GENOVA, a suite of programs that are generally considered to represent the gold standard for generalizability analysis and placed them in the public domain. Included is urGENOVA, a command line program running in DOS, which calculates univariate, random effect variance components for moderately unbalanced datasets employing the analogous ANOVA procedure (Henderson's method 1).

However, using urGENOVA is not for the faint of heart. While it performs Steps 3–6 very efficiently, it requires a control file with a somewhat picky syntax and does not understand Windows file names. Neither does it perform Step 7 of the generalizability analysis. For these reasons, urGENOVA is not used as widely as it deserves. But duplicating the superb functionality of urGENOVA does not appear to be very rational. We have, therefore, written G_String, a user friendly, visual Windows program as a wrapper around urGENOVA that takes users relatively seam- and painlessly through all seven steps.

Early versions of G_String contained only Steps 1–6, but repeated comments from users motivated us to incorporate Step 7 as well. Originally, we employed a series of "IF . . . THEN . . . ELSE . . ." tests. But in view of the proliferation of experimental models that users wanted to analyze, we now employ transformational syntax. This not only simplifies the program, but it will also make it easier, to incorporate further models in the future.

As shown in "signal and noise" and "two-way ANOVA" sections, the calculation of mean-square differences involves taking a small difference of two large accumulated sums. This type of calculation can lead to large rounding errors. We have considered it advisable, therefore, to renormalize the scores by subtracting the grand mean before summation. This keeps the sums from growing linearly with the number of items summed up. Consequently, rounding errors remain confined. This normalization brings another benefit. G_String ignores empty cells, i.e., missing values in the calculation of the various sums. This means that missing values are automatically replaced by the grand mean of scores. In effect, this solution underestimates the effect due to missing scores.

G_String has three operating modes:

(i) design and number of the various facet levels have to be specified by the user,
(ii) design and respective index columns have to be specified by the user, but the program then determines the levels automatically, and
(iii) the parameters for the G analysis have been stored previously and are to be re-used for a D-analysis.

The program then leads the user gently through Steps 2–7 with a series of dialogs. A title for the dataset and descriptive remarks are stored. Design parameters and location of the data file are solicited and some options can be specified. The program then generates the control file required by urGENOVA. Temporary copies of this control file and the appropriate data file are placed in a special directory containing urGENOVA. The program then hands control over to urGENOVA. After completion of Steps 3–6, urGENOVA generates a result file (extension ".lst") which G_String uses to calculate the ICCs. The results of Step 7 together with an explanation of the logic are copied into the result file as well.

The final results appear both on the screen as well as in the output file. After the G study has been completed, any number of D studies can be performed by entering the number of relevant facet levels and facet modifiers into the appropriate field and clicking on the "D study" button.

Summary

In this section, we have discussed a variety of computer techniques and programs used to calculate generalizability coefficients. We emphasized G_String, a program freely available at McMaster University, because we are most intimately familiar with it.

Worked examples

In this section, we describe a number of common designs ranging from simple, classical one factor reliability designs reformulated in G theory nomenclature, to complex multi-facet nested designs. We have obviously not exhausted the possibilities, but rather have attempted to identify and provide examples of some of the more common designs.

The intent is to demonstrate how each design is formulated in the notation of G_String. We describe each design then reformulate the design in G theory language. We describe any specific requirements for the format of the input data, and the sequence of inputs to the screens required to specify the design. Finally, we show how to iterate values on Screen 12 to conduct a variety of D studies.

One-facet designs

DESIGN 1.1. Inter-rater reliability

A clinical researcher examines clinician judgment of severity of illness for patients with congestive heart failure. She locates complete records of 75 patients, and distributes these to three respiratory physicians, who rate each case on a 0–100 scale, where 100 is "Perfect Health."

This example is a typical design for CTT. However, for illustrative purposes, we will recast it as a G theory study. The facet of differentiation is patient; the single facet of generalization is rater. The design is crossed.

The input screens would resemble:

			Design	
Step 3	Subj. Population	Abbrev.	Crossed	Nested
	Patient	p	•	

Step 4	Number of facets	1

			Design	
Step 5	Facet name	Abbrev.	Crossed	Nested
	Raters	r	•	

Step 8	p	75
	r	3

The G study output automatically generated on Screen 12 would look like:

			Generalized across		
Step 12	Facet name	Different	Random	Fixed	Levels
	Patient	•			75
	Rater		•		3

Note that the computed G coefficient is for the average of all raters. To calculate inter-rater reliability for a single rater, you enter "1" as levels for rater, and rerun.

			Generalized across		
Step 12	Facet name	Different	Random	Fixed	Levels
	Patient	•			75
	Rater		•		1

GENERAL TIP:

Often people distinguish between agreement on nominal variables, which should be analyzed with Kappa or weighted Kappa, and reliability with measured variables, which can be analyzed with ANOVA methods and intra-class correlations. However, Fleiss and Cohen (Fleiss & Cohen 1973) showed the two methods are mathematically identical. This means that you can use the power of G theory even with data like 1 = dead, 2 = Alive, see Streiner and Norman (2008b, Health Measurement Scales, 4th ed., pp. 187–188)

DESIGN 1.2. Questionnaire

The researcher administers a questionnaire on "learning style" with 25 questions and "Strongly Agree" → "Strongly Disagree" and seven-point scales to a sample of first year medical students ($n = 125$). He analyses the data to calculate the internal consistency reliability (Cronbach's α)

Again, this can be handled with CTT; however, we will cast it in G theory framework. The facet of differentiation is "student" (s) with 125 levels and the facet of generalization is "item" (i) with 25 levels. Typically the data would be laid out on a spreadsheet with 125 lines, and 25 columns. Input screens would look like:

			Design	
Step 3	Subj. Population	Abbrev.	Crossed	Nested
	Student	s	•	

Step 4	Number of facets	1

			Design	
Step 5	Facet name	Abbrev.	Crossed	Nested
	Items	I	•	

Step 8	s	125
	I	25

The design is formally equivalent to the previous design. The G study output automatically generated on Screen 12 would look like:

			Generalized across		
Step 12	Facet name	Different	Random	Fixed	Levels
	Student	•			125
	Item		•		25

However, in this case, no further analysis is necessary. Internal consistency is the reliability of the average score or total score across all items (Streiner & Norman 2008a, pp. 88–93) which is the G coefficient computed automatically. We could then do D studies varying number of items to determine the effect *n* reliability.

DESIGN 1.3. Teacher rating

A researcher examines the reliability of teacher ratings. The analysis is based on the total score over five items, with five-point scale "Agree" → "Disagree" responses. There are five teachers involved in the study, with each teacher responsible for a different section. Varying numbers of students completing the ratings – teacher 1 – 12 students; teacher 2 – 17 students; teacher 3 – 9 students; teacher 4 – 15 students; teacher 5 – 22 students.

This design introduces a new concept – *nested* facets. Student (s) is nested in teacher (t); since each student can appear with only one teacher. The design is also *unbalanced* – different numbers of students per teacher.

In laying out the data, it is important to note that, while each row in the spreadsheet will likely contain the five ratings of each student, in contrast to the previous examples, the facet of differentiation is not equivalent to the row. We are differentiating teachers, and student now is a *rater* of the teacher, so student is the facet of generalization. Because G_String identifies data by location in the database, not identifier, all records for each teacher must appear in sequence in the database.

The input screens would now look like:

			Design	
Step 3	Subj. Population	Abbrev.	Crossed	Nested
	Teacher	t	•	

Step 4	Number of facets	1

			Design	
Step 5	Facet name	Abbrev.	Crossed	Nested
	Student	s		•

Step 6: You declare the nature of the nesting in Screen 6, by dragging "s" (on the left) to "t" (on the right)

Step 8	t	5				
	s	12	17	9	15	22

Note the differing number of levels for student at each level of teacher The G study output automatically generated on Screen 12 would look like:

			Generalized across		
Step 12	Facet name	Different	Random	Fixed	Levels
	Teacher	•			5
	Student : Teacher		•		13.7

Note the fractional number of levels of student. This is because the harmonic mean is used for these calculations (see p. 28). You can proceed to do D studies, to determine the relation between number of raters and reliability by simply overwriting the "levels" in student and recalculating.

Two-facet designs

DESIGN 2.1. Raters and items

To examine the reliability of the abstract review process for a recent conference, the chair assembled 30 abstracts at random, and had five judges rate each abstract on four items – creativity, methodological rigor, analysis, practical relevance, each with five-point poor → excellent scales.

This is a straightforward two facet, crossed design. However, it is critical to recognize that the "object of measurement" is not a person (the rater) but the abstract. The data must be laid out with raters grouped within abstracts – that is, Abs 1 – Rater 1, Abs1 – Rater 2, Abs 1 – Rater 3, Abs1 – Rater 4, Abs1 – Rater 5, Abs2 – Rater 1, Abs 2 – Rater 2, Abs2 – Rater 3, etc. These may occur on the same or separate lines (which is handled in Screen 7) but must occur in this sequence.

The input screens would now look like:

			Design	
Step 3	Subj. Population	Abbrev.	Crossed	Nested
	Abstract	a	•	

Step 4	Number of facets	2

			Design	
Step 5	Facet name	Abbrev.	Crossed	Nested
	Rater	r	•	
	Item	I	•	

Step 8	a	30
	r	5
	I	4

The G study output automatically generated on Screen 12 would look like:

			Generalized across		
Step 12	Facet name	Different	Random	Fixed	Levels
	Abstract	•			30
	Rater		•		5
	Item		•		4

G_String automatically computes the G coefficient corresponding to the average score over five raters and four items (dividing error variances by 5, 4, or 20). You can also modify this screen to calculate the G theory equivalent of inter-rater reliability and internal consistency (α). To do this, the general strategy is to set the facet of interest as a random facet and set the other facets as fixed facets. You then modify the number of levels of the facets. The basic idea is that the number of levels of each facet is the number of observations that will be used to average the error variance, either of random or fixed facets.

Thus, if you wish to examine inter-rater reliability, "*i*" is set as fixed. Then the number of levels of "*r*" is set to one, since, as described in Example 1.1, you want to compute the reliability of a single rater. If you want to compute inter-rater reliability of the total score, no. of levels of "*i*" remains at four; if you want to compute the inter-rater reliability for a single rating, "*i*" is set to one. The possibilities, then, are:

Inter-rater – one item:

			Generalized across		
Step 12	Facet name	Different	Random	Fixed	Levels
	Abstract	•			30
	Rater		•		1
	Item			•	1

Inter-rater – average score:

			Generalized across		
Step 12	Facet name	Different	Random	Fixed	Levels
	Abstract	•			30
	Rater		•		1
	Item			•	4

Internal consistency (α):

			Generalized across		
Step 12	Facet name	Different	Random	Fixed	Levels
	Abstract	•			30
	Rater			•	1
	Item		•		4

Average inter-item correlation:

			Generalized across		
Step 12	Facet name	Different	Random	Fixed	Levels
	Abstract	•			30
	Rater			•	1
	Item		•		1

GENERAL TIP

It is always important to be very careful in determining which facet represents the "object of measurement" or equivalently, the facet of differentiation. As in the example above, it is not always the people who are completing the questionnaire. Serious errors can result. Further, the data may be analyzed with different facets of generalization, depending on the question (see Streiner & Norman 2008b, p. 241, for an example).

DESIGN 2.2. Questionnaire with multiple subscales

A researcher assesses quality of life for a cohort of patients ($n = 50$) with multiple sclerosis using a quality of life scale with three subscales – physical – 20 items; social – 12 items; emotional – 7 items. She examines internal consistency from the single administration.

The study is quite common. Essentially, from the single administration, you can examine internal consistency within scale and between scales. The facet of differentiation is "patient" (*p*) with 50 levels; there are two facets of generalization: subscale (*s*) (3 levels) and item nested in subscale (i : s), (20, 12, and 7 levels). The data would typically have one line perpatient, with 39 observations on each. Input would look like:

			Design	
Step 3	Subj. Population	Abbrev.	Crossed	Nested
	Patient	p	•	

Step 4	Number of facets	2

			Design	
Step 5	Facet name	Abbrev.	Crossed	Nested
	Scale	s	•	
	Item	I		•

Step 6: Drag "*I*" from left to "*s*" on right.

Step 8	P	50		
	s	3		
	I	20	12	7

The G study output automatically generated on Screen 12 would look like:

			Generalized across		
Step 12	Facet name	Different	Random	Fixed	Levels
	Patient	•			50
	Scale		•		2.6
	Item : Scale		•		13

Note the unusual number of levels for both scale and item: scale. These formulae are described on pp. 33–34.

The G coefficient represents the internal consistency of the overall scale consisting of the three subscales with variable number of items. You can then compute various other combinations, similar to the D study manipulations in the previous example.

(1) Generalizability across scales

Set scale random, item fixed. Set number of levels for scale = 1, leave items: scale at 13. This then computes the average correlation between scale scores.

(2) Generalizability across items within scale

Set scale fixed, item: scale random. Set number of levels for scale = 1, leave items: scale at 13. This then is the *average* internal consistency within each subscale.

However, generally, one would report the internal consistency of each scale individually since the number of items and the specific items vary across scales.

To do this, you would do separate runs for each subscale, using item as the only facet of generalization, as in Design 1.2, and using the feature of Screen 9 to change the starting point.

(3) Overall internal consistency, independent of subscales

Simply rerun as Design 1.2, with item having 39 levels.

Note that it is difficult to compare α's derived from different scales as α is sensitive to the number of items in the scale.

Multiple facet designs

The introduction of additional facets involves additional complexity, but no new concepts. The critical steps are to first identify object of measurement, then label the various additional facets in the design, identify which are nested and which are crossed, and then ensure that the sequence of data in the spreadsheet lines up with the intended design.

Stratification facet designs

One other class of designs that is very common in generalizability studies in medical education. Particularly for performance tests like OSCEs and oral examinations, it is very common to run the examination at multiple sites over several days. In these circumstances, each subject can be said to be nested in a particular "stratum" of a stratification facet (day, site). To complicate things further, it is very common to change raters, or in the case of OSCEs, to also change the specific stations to ensure test security. Thus, both participant (*p*) and possibly station and rater are nested in one or more "stratification" variables – site, day, circuit.

GENERAL TIP

G_String and urGENOVA are not currently capable of dealing with designs when a facet of generalization is nested in a stratification facet. The next version (G_String V) will have this capacity.

DESIGN 4.1. You are running an OSCE which is taking place in two different hospitals. Students (*p*) are randomly assigned to one hospital or the other. At each hospital the same 12 stations are used. Three circuits are run at hospital A; for a total of 36 students and 4 circuits at hospital B, for a total of 48 students. Each station has a station – specific checklist with anywhere from 12 to 27 items.

This is a very typical OSCE setup identifying the facets from slowest (supraordinate) to fastest (subordinate). The first

stratification variable is hospital (h) with 2 levels, then circuit : hospital (c : h) with three and four levels. Then participant : circuit : hospital (p : c : h). Crossed with this is station (s) and item : station (i : s).

Data need to be laid out consistently with this hierarchy, likely with one physical record per applicant or per station. As before, caution must be exercised to ensure that the records are grouped according to this hierarchy.

The screens will now look like:

			Design	
Step 3	Subj. Population	Abbrev.	Crossed	Nested
	Participant	p		•

Step 4	Number of facets	4

			Design	
Step 5	Facet name	Abbrev.	Crossed	Nested
	Hospital	h	•	
	Circuit	c		•
	Station	s	•	
	Item	I		•

Step 6: Drag "c" to "h", "p" to "c : h" and "i" to "s".

Step 8	h	2											
	c : h	3	4										
	p : c : h	12	12	12	12	12	12	12	12	12	12	12	12
	s	12											
	i : s	14	22	17	...	....	...	...	..	..	..	..	..

The G study output automatically generated on Screen 12 would look like:

			Generalized across		
Step 12	Facet name	Different	Random	Fixed	Levels
	Participant per circuit	•			12
	Hospital				2
	Circuit per hospit				3.4
	Station		•		10.7
	Item per station		•		18.5

Note that (a) hospital and circuit do not have an dot. This signifies that they are stratification facets. (b) The number of levels for station and item : station contain fractions, which reflects the unbalanced design (p. 27).

The resulting G coefficient is the overall test reliability. D studies can be conducted using the strategies discussed previously to examine the average inter-station correlation (S random, I fixed, $n(s) = 10.7$) or the internal consistency among items within station (I random; S fixed $n(i) = 18.5$).

What about the stratification facets? Basically, any variance due to the stratification facet represents a bias, so that one circuit or hospital is, on average, harder or easier than another. The hope or expectation is that these variances will be small. If participants are judged relative to others in the same stratum, this variance is of no consequence, as reflected in the G coefficient for "relative error". However, if absolute interpretation is placed on scores, variance due to strata is a source of error. Therefore, it has to be included in the absolute error" calculation.

DESIGN 4.2. You are running an OSCE which is taking place with residents at two levels. Residents (r) are either PGY1 (36 residents) or PGY4 (48 residents). Residents go through the OSCE 12 at a time, with all residents at each level together. Each station has a station – specific checklist with anywhere from 12 to 27 items.

This design is deliberately set up to be identical in layout to the previous study. The only difference is the meaning attached to one stratification facet. In the previous example, hospital was the supraordinate facet, and the expectation (or hope) was that this would contribute no variance. Any variance due to hospital was treated as error variance which would confound interpretation of scores. Thus the absolute error coefficient best represented the overall generalizability.

In the present case, the expectation is that difference in educational would be large, amounting to a test of construct validity. The statistical test can be easily extracted from the G_String ANOVA table. By including, education in the design, the G coefficient is then determining the ability of the test to differentiate among residents *within* an educational level, which is completely appropriate. This is obtained from the relative e coefficient.

Nested designs

There is one final class of designs that is very common in generalizability studies in medical education. This is the situation where there are multiple and variable numbers of ratings on the object of measurement, with rating nested in the object of measurement ($g : d$) designs. One example is teacher ratings, where students in each class rate their teacher. Student is nested in teacher, and number of students will likely vary. Peer assessments of practicing physicians, called "360° evaluation" or "multi-source feedback" is another – different peers with different numbers of observations for each physician. Typically these are not the only two facets, since often ratings are on multi-item questionnaires, so the design would be peer nested in doctor crossed with item. Another common variant is the so-called "mini-CEX" where each student is observed

on a number of occasions by her supervisor(s), and again, typically each student has different supervisors.

Frequently these designs can have very many observations. One study involved over 1000 physicians rated by 17,000 peers. Another was based on a teacher evaluation system at a large university and had 65,000 observations on 1700 teachers. To handle these studies in previous versions of G_String is very tedious as the number of observations in each nest had to be entered manually. However, with G IV, all one need do is assign a unique index to each teacher or physician and another unique index to each rater, creating two column variables. G IV will read these indices and automatically create the correct number of levels in each nest.

There is one common variant of this design. Frequently the same rater may be involved in multiple ratings of the subject. For example, with students in community clinical rotations, each student may receive multiple observations and ratings from the same rater. This is handled in G_String simply by creating a third "sequence" index which is unique for each rating, so that the design becomes $g_2 : g_1 : d$ (sequence : rater : student).

While this design can be analyzed, *extreme caution must be exercised in interpretation*. The problem is that with multiple ratings from each rater, rater variance (lenient – stringent) is confounded with subject variance. In the extreme case, where each subject is rated by one rater, different for each subject, rater and subject variance are completely confounded. One can obtain high G coefficients, but the value is biased upwards since this results from variance due to rater and variance due to subjects.

As a heuristic rule, G_String issues a cautionary message if the average number of (nested) raters per subject is less than three.

GENERAL TIP

With designs where facets of generalization (raters) are nested in facet of differentiation, exercise extreme caution in situations where there are multiple observations from individual raters.

DESIGN 5.1. You are collecting data from your undergraduate program to assess teacher effectiveness. You have seven undergraduate courses, with numbers of students varying from 12 to 145. Although this is not strictly true, assume in this example that students are different in each course. These ratings are done after random lecture, so ratings are available for varying numbers of lectures per teacher. The form has 11 items.

This is a $g_3 \times g_1 : g_2 : d$ study, where the facet of differentiation is teacher, the facets of generalizations are lecture, student, and item. Typically, there would be one physical record for each rating with 11 ratings. To analyze in G IV, the ratings should be identified with three indices – teacher, lecture, and student, in that sequence. Data must be sorted in ascending order on each of these indices.

The screens will now look like:

			Design	
Step 3	Subj. Population	Abbrev.	Crossed	Nested
	Teacher	t	•	

Step 4	Number of facets	3

			Design	
Step 5	Facet name	Abbrev.	Crossed	Nested
	Lecture	l		•
	Student	s		•
	Item	I	•	

At Steps 4 and 5, a "column" box will also appear on the right. You will indicate in what column on the record the index for teacher, lecture, and student is located. For the item facet, which is multiple observations on each record, you can either leave column blank and enter number of items at Step 8, or insert "−1" and G IV will compute the number of items. If there are items within scales, on the same record, you can simply enter the number of levels of each at Step 8.

Step 6: Drag "l" to "t", "s" to "l : t".

At Step 8, G IV will automatically generate the number of levels for t, l, and s (and I if column is −1)

The G study output automatically generated on Screen 12 would look like:

		Generalized across			
Step 12	Facet name	Different	Random	Fixed	Levels
	Teacher	•			7
	Lecture Teacher		•		3.2
	Student: Lecture : Teacher		•		17.9
	Item		•		11

Note that the number of levels for lecture and student contain fractions, which reflects the unbalanced nested design (p. 27).

Caution

Once again, we emphasize the potential for bias in the design as a result of confounding between rater and teacher (*g* facet and *d* facet). If, for example, ratings of all lectures for each teacher were done by a single paid student in the class, then rater variance is confounded with teacher variance and coefficients are interpretable.

Summary

In this section, we worked through a series of real-world examples, typical for what one commonly encounters in educational practice.

Conclusion

In this Guide, we have tried to introduce G theory through the use of a step-wise practical approach, using examples commonly found in the world of medical education. We have no doubt that many readers find the theory difficult to master but with the help of this guide, plus G_String, a program specifically designed to aid researchers in the use of G theory and some worked examples in the appendices, we hope that the subject is more clear. They say "practice makes perfect;" we hope through this Guide you "practice" G theory and we have been of help in assisting you in mastering the subject.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Notes on contributors

RALPH BLOCH, is retired and lives for his family and hobbies. Previously, he was Professor of Medical Education and Director of the Institute for Medical Education at the University of Berne, Switzerland, Professor of Rehabilitation Medicine at McMaster University and part-time Professor in the Department of Clinical Epidemiology and Biostatistics at McMaster University in Hamilton, ON. He has published 65 articles and 4 books.

GEOFFREY NORMAN, PhD, is a Professor of Clinical Epidemiology and Biostatistics, McMaster University and Assistant Dean of the Program for Educational Research and Development at McMaster. He has won numerous awards and published about 250 articles and 10 books.

Notes

1. The authors, both amateur cabinet makers, adhere to the maxim: measure twice, cut once.
2. Cardinet (1975), Tourneur and Allal were the first to point out that the "object of measurement" may change, and can be viewed as one more source of variance.
3. We call this a "three facet" design, referring to the number of facets of generalization.
4. This rule indicates that the definition of a stratification facet is that the object of measurement (Person) is nested in it.
5. The term "stratification" is consistent with the terminology of Brennan, 2001, Section 5.2.

References

Allal L, Cardinet J. 1976. Application of generalizability theory: Estimation of errors and adaptation of measurement designs. Neuchâtel: Institut Romand de Recherches et de documentation pédagogiques.

Brennan R. 2001. Generalizability theory. New York: Springer Verlag.

Cardinet J. 1975. The generalizability of surveys of education outcomes. Neuchatel, Switzerland: Institut Romand de recherches et de documentation pedagogiques.

Cardinet J. 2010. EduG. Edumétrie: Qualité de la mesure en éducation. [Accessed 30 June 2012] Available from http://www.irdp.ch/edumetrie/precedent_e.htm

Cohen J. 1960. A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46.

Cronbach L, Gleser G, Nanda H, Rajaratnam N. 1972. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Cronbach L, Nageswari R, Gleser G. 1963. Theory of generalizability: A liberation of reliability theory. Br J Stat Psychol 16:137–163.

Degrees of Freedom. (n.d.). [Accessed 30 June 2012] Available from http://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics)

Fisher RA. 1925. Intraclass correlation and the analysis of variance. In: Statistical methods for research workers. New Delhi: Cosmo Publications for Genesis Pub., 2006.

Fleiss J, Cohen J. 1973. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 33:613–619.

Keller LA, Clauser BE, Swanson DB. 2010. Using multivariate generalizability theory to assess the effect of content stratification on the reliability of a performance assessment. Adv Health Sci Educ Theory Pract 15(5):717–733.

National Institute of Standards and Technology. (n.d.). NIST statistics portal. [Accessed 20 February 2012] Available from http://www.nist.gov/statistics-portal.cfm

Norman G. 2010. Likert scales, levels of measurement and the "laws" of statistics. Adv Health Sci Educ Theory Pract 15:625–632.

Pearson K. 1896. Mathematical contributions to the theory of evolution III. Regression, heredity and panmixia. Philos Trans R Soc Lond Ser A 187:253–318.

Shavelson R, Webb N. 1991. Generalizability theory: A primer. Thousand Oaks, CA: Sage.

Streiner D, Norman G. 2008. Biostatistics: The bare essentials 3/e (with SPSS). Hamilton, ON: BC Decker.

Streiner D, Norman G. 2008. Health measurement scales: A practical guide to their development and use. 4th ed. Oxford: Oxford University Press.

University of IOWA. (n.d.). Center for advanced studies in measurement and assessment. Iowa City, IA: The University of IOWA, College of Education. [Accessed 11 July 2012] Available from http://www.education.uiowa.edu/centers/casma/computer-programs.aspx#genova

Winer J, Brown DR, Michels KM. 1991. Statistical principles in experimental design. New York: McGraw Hill.

Appendix A

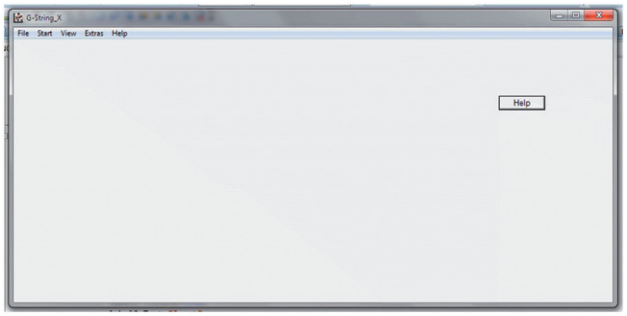
Getting started with G_String

G_String guides the user through all the steps of setting up a control file for urGENOVA, feeds the control file to urGENOVA, and allows the user to inspect and modify the control file and view the result file via a familiar Windows® user interface. G_String has built-in *help screens*. After urGENOVA has executed, G_String can then compute G coefficients under user control.

To start G_String, click on G_String.exe or a shortcut. Then, in G_String click on "Start."

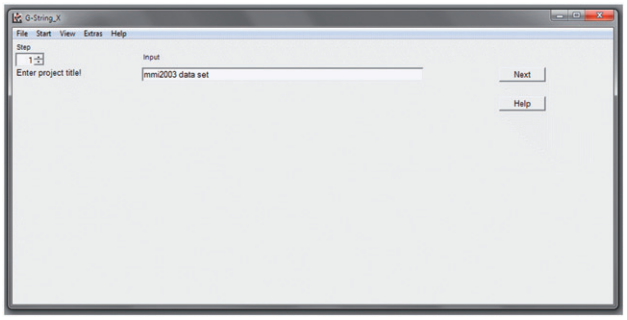
At this point, a submenu with three options is displayed. "*Start fresh*" is the usual approach, where you are creating a new G_String run and all facets and all levels of each facet will be user-specified. "*Start over*" enables you to do multiple runs of the same database, in order to perform or refine D studies that were not done during the initial analysis. Selecting "*Auto index*" tells G_String to automatically count the number of levels of each facet. As described in detail see subsection 'Specifying the number

of levels', this is very useful for unbalanced nested designs with large numbers of subjects and/or raters.



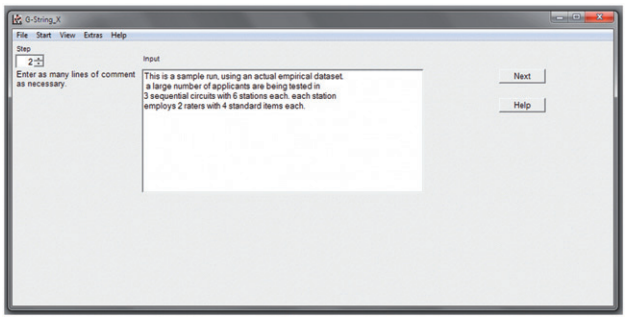
Step 1: Selecting a title

"Title" can be any combination of letters and numbers up to 80 characters. It is not actually used in the calculations, so can be omitted, but it appears in the output.



Step 2: Entering comments

Comment fields are optional and are not used in the calculations but copied into the result (output) file. G_String adds some comment lines automatically.



Step 3: Defining "subjects"

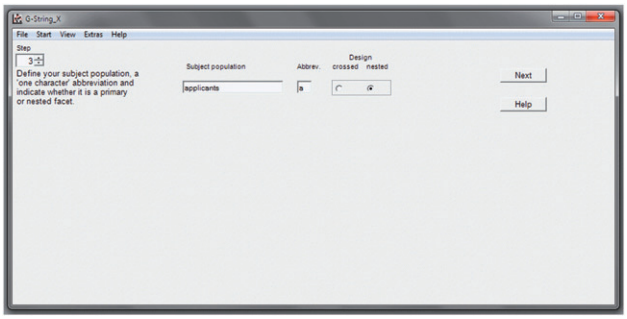
"Subject" is the variable describing the people or things that were measured in the study – the "object of measurement." This is also the "facet of differentiation." In Brennan's terminology, "subject"' is always labeled *p*. While in G theory the designation "subjects" is to some extent arbitrary, usually reliability or G coefficients are referenced to subjects. Usually, but not always, the data records are arranged subject by subject.

"Subject" is usually crossed with other factors, such as item or rater (e.g., a series of students being rated by three raters on a 10-item test), which would be the repeated measures in a simple analysis. However, "subject" may also be nested.

Example: Student may be nested in year (freshman, sophomore, senior); patient may be nested in gender or physician practice, and can be both crossed with some variables and nested with others. G_String easily deals with this situation. Facets such as year, gender, physician as above are labeled "stratification facets" and are handled somewhat differently, as will be described (Brennan 2001, p. 153).

While, in principle, "subject" may be nested in many stratification facets, in practice G_String is restricted to four stratification facets.

If "*auto-index*" is selected, a column box will also be displayed. You must specify in which column of the database the index for the "subject" facet is located; this is described earlier.



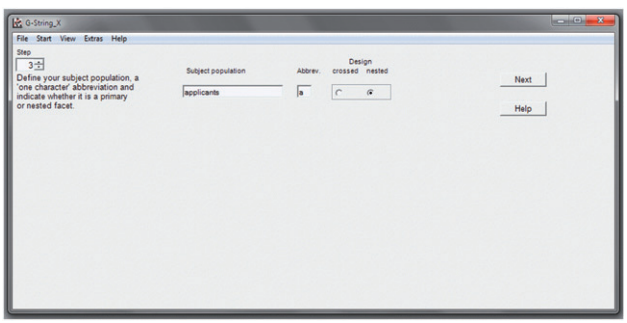
Step 4: Defining the other facets

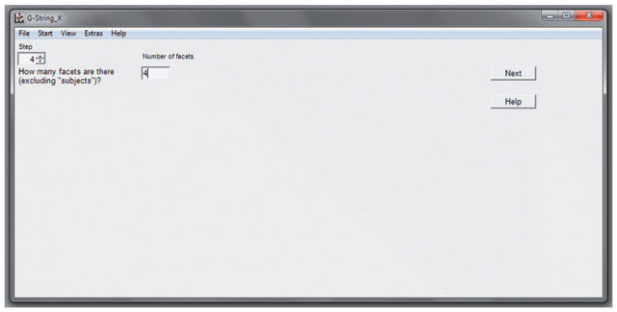
A "facet" in a design is any factor (in ANOVA jargon) or variable used to categorize the data for analysis. In G theory, "subject" is always a factor, and is not counted explicitly at this step. Some variables are crossed with others, some are nested.

Example: The present example is a six-station OSCE. There were three circuits (C), with six applicants (A) each. Applicant is nested in circuit. Station is crossed with applicant (all applicants do all stations). All stations have two raters, with the same four items in each station Therefore, item is crossed with station but rater is nested in station, since each station has its own raters but items were constant across stations.

In Step 4, you simply specify the number of facets in addition to subjects. For the OSCE, this would be four (circuit, station, item, rater).

As described earlier, any number of facets with fixed levels occurring on the actual record line can be specified. For this purpose you leave the column fields empty for these facets. You will then be prompted to manually enter the actual fixed levels. If, however, the numbers of levels per record line have to be determined automatically, the record line may contain only one facet. In this case enter "−1" in the corresponding column field.





Step 5: Naming and specifying the facets

In this step, you name the facets and indicate which are nested in other facets.

- Give each facet a descriptive name and a corresponding one-character, unique, lowercase abbreviation.
- If a variable is nested in one or more other variables (see Step 4), then you change the default "crossed" to "nested."
- In the OSCE example, applicant is nested in circuit, (Screen 3) and rater is nested in station.
- Variables must be listed in the order they are encountered in the data file, from slowest moving to fastest.

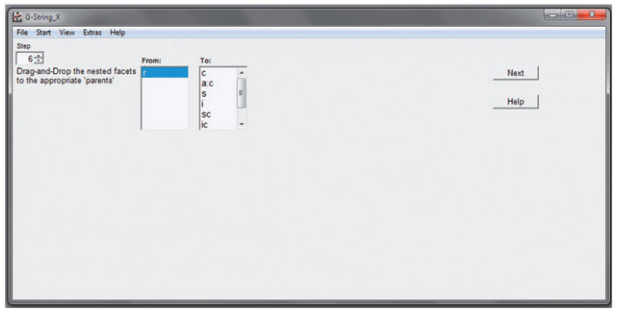
In the OSCE example: if the data have one record per student, with all data for each station, then the data for each rater, then the responses on each item, the order of additional variables would be: circuit, station, rater, item.

Step 6: Facet nesting

In this step, nested factors are "drag-and-dropped" to the right side so that they are located under the factor in which they are nested. Every possible combination of crossed facets is shown in this box, and a facet can be nested in more than one other facet, e.g., a : ic.

Pick a nested facet up with the mouse cursor from the list on the left and drop it on the desired combination in the list on the right.

In the example, applicant has already been dragged under circuit. Rater will be dragged to station (s)



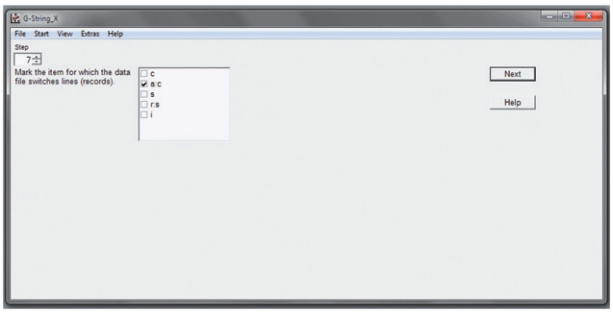
Step 7: Identifying the data structure

Based on the specification of nested and crossed factors in Step 6, G_String creates a list corresponding to the order in which the data are expected to occur.

In the OSCE example, the list would be: subject then station then rater then item, listed as:

- circuit,
- applicant : circuit,
- station,
- rater : station,
- item.

You now specify which variable corresponds to the physical record (in Excel, each row). For example, if all data for one student was on one line, the check is put beside "applicant : circuit" (a:c). If each station is listed on one line (with all raters and items), the check is beside station.



Step 8: Specifying sample sizes

(*If "auto index" is selected, the number of levels of each facet will be computed automatically and the corresponding fields will contain the appropriate number of levels. When the number of detected levels is more than 30, their value will not be displayed.*)

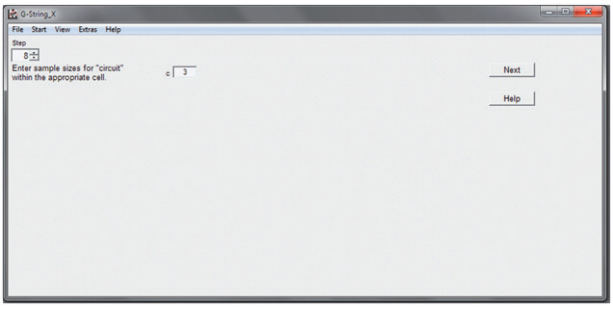
At this step, G_String cycles through all the variables you specified, and asks for "sample size." The "sample size" is the number of levels of each facet and must be >1.

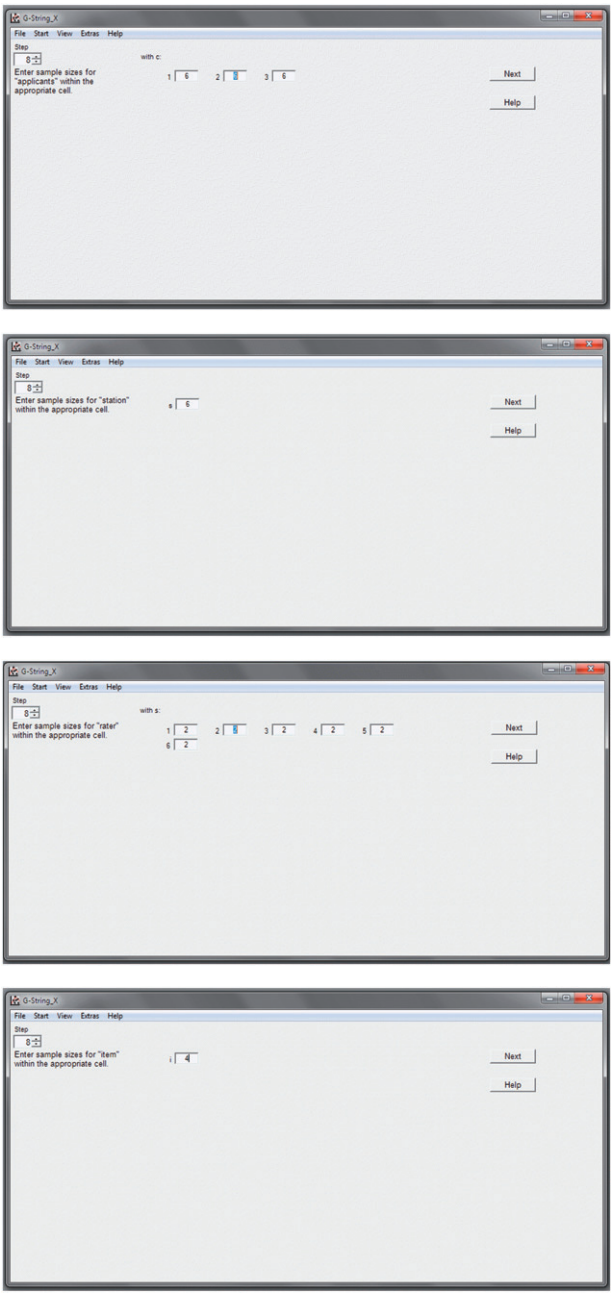
In the OSCE example, "sample size" for station is just the number of stations.

For *nested* variables, you must specify the number of levels at each level of the nesting variable.

For subject, this is the number of applicants in each circuit 6, 6, 6. For rater this will be the number of raters per station, 2, 2, 2, 2, 2, 2.

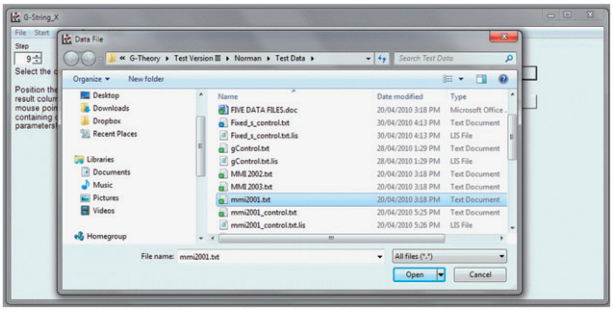
As a default, once you enter the levels for the first box and press the "tab" key, G_String will automatically assign the same number of levels for all boxes. If the numbers differ, simply overwrite the pre-assigned numbers. The sequence below illustrates how all the levels are being entered.





Step 9: Locating and specifying data file

At this step, you first tell G_String where the data file is located using the usual Browse function. G_String then reads the first few records from this file. It assumes that the actual data are listed sequentially beginning at a specific column of each data line in the data file. Recall that data must be in an ASCII text file.

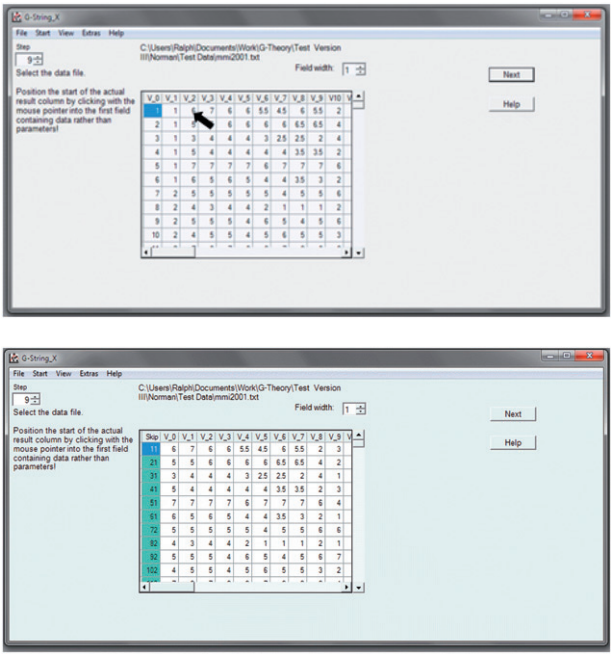


You then select the column where the actual data start by mouse-clicking directly on the first cell containing data (in any row). urGENOVA will ignore anything to the left of this.

For tab-delimited files, G_String will create the correct field width. For fixed field data (no delimiters), first set the start column as above, then with the "field width" selector, indicate the width in columns of each individual data field (including blanks).

In the example, the first two columns are identifiers, so the cursor is placed in the third column.

The cursor arrow must be located in the first actual data field, *not* on the headers.



Step 10: Options

urGENOVA allows you to specify a number of options. G_String assumes some default values that you do not have to change, unless you know what you are doing.

NREC: the number of data records that will be printed in the output file. Useful to check that the data are being read as expected.

Outname: the name of the output file. This will be assigned a name and stored in the same directory as the data file, unless you choose a new name and directory.

ET prints the expected T-term equations.

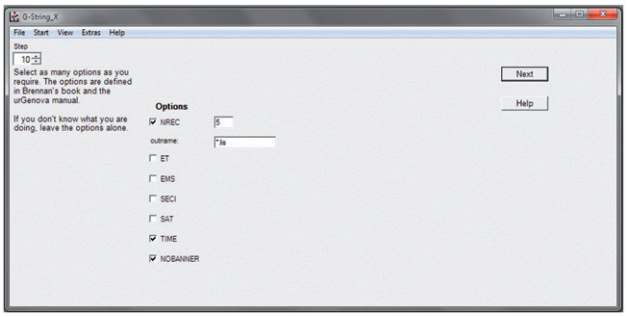
EMS prints the equations for the expected mean squares as sums of variances.

SECI .nn is the standard error and ".nn" confidence interval for the estimated variance component (.nn is a fraction between 0.00 and 1.0, usually 0.95).

SAT is a second confidence interval estimate, due to Satterthwaite (see the GENOVA manual).

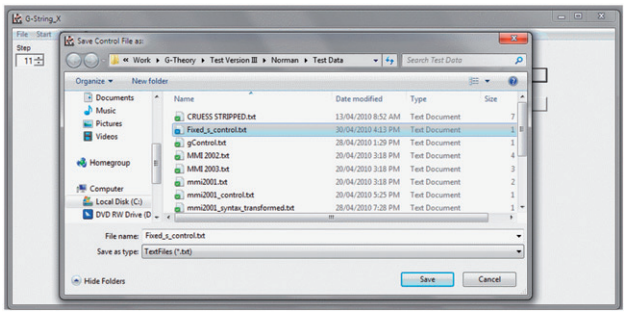
TIME: Time and date of processing will be printed (default is ON).

NOBANNER Banner will not be printed (default is ON).



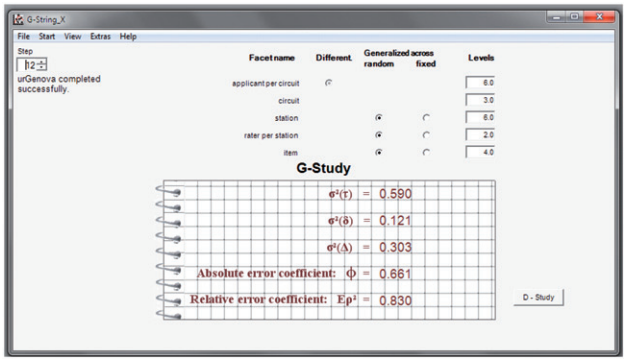
Step 11: Save control file

You have now completed the specification and generated a control language file. By default, it is called "gControl.txt" and stored in the same directory as the data file; however, at this step you can give it a more meaningful name and place it in any directory of your choice.

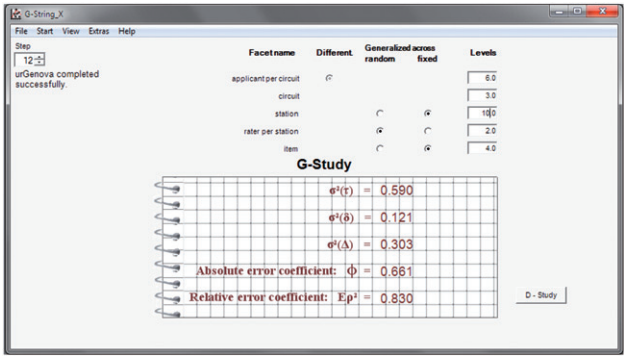


Step 12: Calculating variance components

Once you saved the proper control file path, urGENOVA is executed automatically to calculate the variance components and the coefficients of variance for the G study are generated.



Step 13: Components of variance in G study and D studies



This screen displays the output from the calculation of the G coefficient, and then permits the user to conduct repeated D studies. The output follows the convention of Brennan, in particular the rules for calculation of G coefficients (4.1.6, p. 109) and the section on Mixed Models (4.3, p. 120). A brief explanation is required.

Generalizability theory is an extension of CTT. In CTT, every observation is comprised of a true score or signal, and error. The reliability coefficient is the ratio of the true score var(τ) to the total variance (var(τ) + var (δ)). G theory extends this formulation by considering that error may have multiple sources, which we have called "facets of generalization." Depending on the measurement situation, you may wish to generalize over some facets (called "random" facets by Brennan), and keep others constant (called "fixed" facets by Brennan).

In the OSCE, if we set rater as random and item and station as fixed, we will compute the equivalent of the inter-rater reliability. If we set item as random and fix rater and station, we are computing the equivalent of internal consistency.

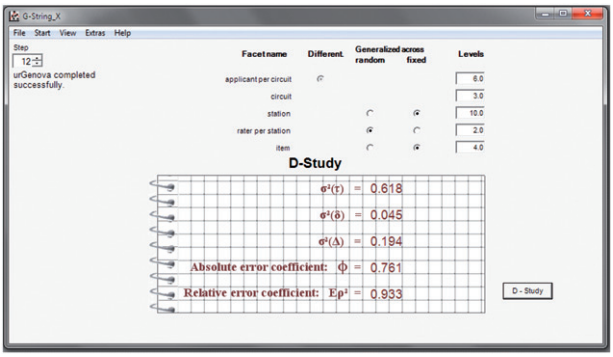
The calculation amounts to moving variance components between the error term δ and the signal term τ. Screen 11 displays δ and τ as well as Δ, described next.

There is a further refinement in G theory. Sometimes, we wish to interpret a person's score *relative* to those of other people. In this situation, the fact that some raters may be more strict or lenient than others, or some items harder or easier, is irrelevant. This amounts to ignoring the main effects of the facets of generalization, and only interactions with subject are included. This is the error term δ. However, if we wish to put an absolute interpretation on scores, we must include main effects, which is the term Δ on Screen 12. In turn, the absolute error coefficient or Φ contains Δ whereas the relative error coefficient or $E\rho^2$ contains δ, see earlier description for further explanation.

The first automatic output on this screen considers all facets as facets of generalization. Further, it computes averages over each facet, based on the sample sizes in the original study. So the calculated coefficient $E\rho^2$ is the G coefficient for the original test.

However, on Screen 11, G_String will calculate G coefficients with any combination of fixed facets and facets of generalization, and any sample sizes – so-called D studies – in order to examine the effect of each facet on the overall generalizability. You can also calculate the equivalent of classical coefficients by "treating" one facet at a time as "random" and fixing the remaining facets.

In the OSCE example, if you want to compute the equivalent of inter-rater reliability in the OSCE, you would: (a) set item and station as fixed facets and (b) set the sample size for rater = 1. (If you keep sample sizes for item and station, you are calculating inter-rater for the average of $N_s = 6$ stations and $N_i = 4$ items.) More likely you would also fix sample sizes for item and station at one to determine inter-rater reliability for a single rater in one station with one item.



If you wish to calculate different coefficients (D studies), simply re-enter the new combination of facets, identifying fixed facets and facets of generalization, and the new sample sizes and click on "compute." The new coefficient will be calculated and displayed in the screen and in the printout. Refer to earlier pages and examples in Appendix B for more detailed explanations.

Note that, in the case of nested variables, the number of levels is within each nest. For D studies you must keep this constant across nests so it is a "balanced" design.

In the OSCE study, there are six applicants per circuit and two raters per station.

Appendix B

Data

Outline

- Data structure
- Data formats

Means, mean-square differences, variances, and variance components are calculated from actual scores corresponding to the specific experimental design. The input to G_String, therefore, is a data-file structured so that the program can accurately allocate each data item to the appropriate subject and facet level. This requirement sets relatively strict conditions on the format of the data-file. Any violation of these restrictions can make the file unreadable or cause errors in the results.

The data-file is an ASCII text file organized in lines (rows) and fields (columns). Data have to begin on the first line. It may not contain title or header lines. Each field is either empty or contains one score value as positive or negative, decimal number. Lines are separated by otherwise invisible "character return" and "line feed" characters. Fields can be defined either as "fixed width" or "delimited". The otherwise invisible "tab" character delimits a field. Using "tab" delimited fields makes the data file more resistant to misreading.

The actual scores do not have to start at the beginning of each line. The format allows for a fixed number of leading characters or columns. These can either contain facet index information or they can be skipped. In nested facets, missing data are meaningless. In crossed facets, missing data should be indicated by delimited blanks.

The easiest way to construct a data file for G_String is to first assemble the data as a spreadsheet file, visually inspect, and edit it, and finally export or save it as a "tab delimited text file (.txt)".

It is highly advisable to start the analysis of a new dataset by first creating a new directory with a unique and descriptive name and place all corresponding data-, control-, and result-files in this directory.

Appendix C

Interpreting program output

The computer output contains many more details of the above calculations and will be described next. This output is generated when the process of study calculation is finished, and is created as a ".txt.lis" file in the target directory. Below is a sample output from the example. Annotations are in this font. On some computer operating systems you may have to delete the secondary extension ".lis" to be able to read the file.

CONTROL CARDS FOR RUN 1
Control Cards File Name: ~Temp.txt
mmi2003 dataset

```
GSTUDY      mmi2003 dataset
COMMENT
COMMENT     Processing date: 06/06/2010 2:49:22 PM
COMMENT
COMMENT     This is a sample run, using an actual empirical
            dataset
COMMENT     a large number of applicants are being tested in
COMMENT     3 sequential circuits with 6 stations each. each
            station
COMMENT     employs 2 raters with 4 standard items each
COMMENT
COMMENT
COMMENT%    applicant (a)
COMMENT%    circuit (c)
COMMENT%    station (s)
COMMENT%    rater (r)
COMMENT%    item (i)
COMMENT
COMMENT     The calculated "Grand Mean" = 4.4010
COMMENT     G_String III normalizes scores by subtracting
            the Grand Mean from each score
COMMENT
OPTIONS     NREC 5 "*.lis" TIME NOBANNER
EFFECT       c          3
EFFECT      * a:c       6 6 6
EFFECT      s           6
EFFECT      r:s         2 2 2 2 2 2
EFFECT      i           4
FORMAT      30  0
PROCESS     "~Temp.dat"
```

This is an image of the control card input for urGENOVA created by G_String in response to user input. Note how the "EFFECT" lines completely describe the design, with circuits, applicants nested in circuits (6/circuit), stations, rater nested in station (2/station.,) and item. The calculated Grand Mean over all the scores is 4.4010.

INPUT RECORDS FOR RUN 1
mmi2003 dataset

RECORD NUMBER 1

1.599 2.599 1.599 1.599 1.099 0.099 1.599 1.099 −2.401 −1.401

urGENOVA images the data on the first five records. Some are omitted from this example. The grand mean has been subtracted from the actual scores.

MEANS FOR MAIN EFFECTS FOR RUN 1
mmi2003 dataset

Means for c

−0.045 0.528 −0.483

Means for a:c

0.089 −0.078 −0.839 −0.214 0.932 −0.161 0.849 −1.339 1.297 −0.318
0.995 1.682 −0.943 −0.672 0.745 −0.016 −1.130 −0.880

Means for s

0.238 −0.283 −0.102 0.207 0.203 −0.262

Means for r:s

0.418 0.057 −0.887 0.321 0.030 −0.234 0.314 0.099 −0.679 1.085
−0.873 0.349

Means for i

0.205 −0.119 −0.047 −0.040

urGENOVA outputs the means for each variable.

Below is the ANOVA table created by urGENOVA. The format is conventional, except that the right column is "variance component" and is used in the calculation of G coefficients. (Negative variance components are set to zero when computing G coefficients.)

ANOVA TABLE FOR RUN 1
mmi2003 dataset

Effect	df	T	SS	MS	VC
c	2	16882.85677	147.89583	73.94792	0.10650
a:c	15	17393.52604	510.66927	34.04462	0.58973
s	5	16778.13368	43.17274	8.63455	−0.27757
r:s	6	17005.35069	227.21701	37.86950	0.46282
i	3	16747.93634	12.97541	4.32514	0.01591
cs	10	17075.20312	149.17361	14.91736	0.17605
cr:s	12	17351.61458	49.19444	4.09954	0.02040
ci	6	16898.97569	3.14352	0.52392	0.00179
as:c	75	18021.96875	436.09635	5.81462	0.27813
ar:cs	90	18619.43750	321.05729	3.56730	0.82002
ai:c	45	17420.10417	10.45920	0.23243	−0.00642
si	15	16803.64583	12.53675	0.83578	−0.00231
ri:s	18	17044.84722	13.98438	0.77691	0.02484
csi	30	17118.02083	14.16204	0.47207	0.01000
cri:s	36	17420.29167	11.87500	0.32986	0.00711
asi:c	225	18144.87500	69.62934	0.30946	0.01113
ari:cs	270	18845.75000	77.54688	0.28721	0.28721
Mean		16734.96094			
Total	863		2110.78906		

(continued)

Grand mean: 0.

Below is the first output from G_String. It is a calculation of the overall test generalizability, so (a) there are no fixed facets, and (b) the number of levels of each facet corresponds to the original study.

The allocation of individual terms is based on the specification of random or fixed facets. This is according to the rules in "computing G coefficients" section, abstracted from Brennan.

Date and time at beginning of Run 1: Sun Jun 6 14:49:22 2010
Processor time for run: 0 seconds

Computation sequence for G study

"a"	Differentiation	6.00
"c"	Stratification	3.00
"s"	Random	6.00
"r"	Random	2.00
"i"	Random	4.00

Pattern	Var. Comp.	Levels	Signature	Rule
c	0.1065	1	s	Delta only
a:c	0.5897	1	ds	tau only
s	0.0000	(6.0)	r	Delta only
r:s	0.0386	(12.0)	r	Delta only
i	0.0040	(4.0)	r	Delta only
cs	0.0293	(6.0)	r	Delta only
cr:s	0.0017	(12.0)	r	Delta only
ci	0.0004	(4.0)	r	Delta only
as:c	0.0464	(6.0)	dr	Delta and delta
ar:cs	0.0683	(12.0)	dr	Delta and delta
ai:c	0.0000	(4.0)	dr	Delta and delta
si	0.0000	(6.0*4.0)	r	Delta only
ri:s	0.0005	(12.0*4.0)	r	Delta only
csi	0.0004	(6.0*4.0)	r	Delta only
cri:s	0.0001	(12.0*4.0)	r	Delta only
asi:c	0.0005	(6.0*4.0)	dr	Delta and delta
ari:cs	0.0060	(12.0*4.0)	dr	Delta and delta

Results

s2(T)	=0.590
s2(D)	=0.303
s2(d)	=0.121
Er2	=0.830
Phi	=0.661

The first five outputs are shown in Screen 11:
Below is an example of D studies. The user can control two aspects of the computation: (a) which facets are random and which are fixed, and (b) how many levels of each. These are used for different purposes:

Random vs. fixed facets. In G theory, one can compute the equivalent of classical coefficients such as inter-rater reliability, internal consistency, and so on, by restricting the analysis, setting one facet at a time as random, and setting the "*n*" for this facet equal to one.

> In the example above, to compute inter-rater reliability for a single rating and a single station, one would declare *rater* as random, *station*, and *item* as fixed, and set all the levels equal to one. If one wanted the inter-rater reliability of the total score over all four items, number of levels of *item* would remain four. To look at internal consistency (across items) *item* becomes the random facet, *rater*, and *station* fixed, and levels remains at four (since internal consistency is for the total score, so amounts to averaging by number of terms).

The number of levels is a matter of judgment, and is based on whether the reliability is for a single (item, rater, station) or for the mean across all items, raters, and stations. To understand how this works, we have taken the above example and created a number of D study scenarios:

Random facet(s)	Fixed facet(s)	N_{rater}	N_{item}	$N_{station}$	Interpretation
S	R,I	2	4	1	Inter-station reliability of total score from two raters and four items
S	R,I	1	1	1	Inter-station reliability for any single item from any rater
S	R,I	2	4	6	Inter-station reliability for total score from two raters and four items
R	S,I	1	4	6	Inter-rater reliability for total score from four items, six stations
R	S,I	1	4	1	Inter-rater reliability for total score on any station
R	S,I	1	1	1	Inter-rater reliability for any item, any station
I	R,S	1	4	1	Internal consistency (across items) for one rater, one station
I	R,S	1	1	1	Average inter-item correlation
I	R,S	2	1	1	Average inter-item correlation for mean of two raters

Changing levels – D studies. To this point, we have set the number of levels as either the original design number or one, depending on whether we wish to compute reliability for the single item or the number of levels of the facet in the original study. We can also vary the number of items at will, to determine the optimal combination of levels of each facet in the design. In this case, the interest is in the overall test reliability, so there are no fixed facets, but we might vary number of levels at will.

Note that when we proceed with D studies, the design is balanced by definition, since we input the number of levels of each facet as a single number. Thus unbalanced designs only arise in the initial calculation of the G coefficient from the original data.

> For example, are we better to have 6 stations with 2 raters ($Nr=2$, $Ns=6$), or 12 stations with 1 rater ($Nr=1$, $Ns=12$)? What do we gain in going from 12 stations to 18?

```
Computation sequence for D study
''a''   Differentiation   6.00
''c''   Stratification    3.00
''s''   Random           12.00
''r''   Random            1.00
''i''   Random            4.00
-------------------------------------------------------------
Pattern Var. Comp. Levels Signature Rule
-------------------------------------------------------------
c       0.1065    1            s    Delta only
a:c     0.5897    1            ds   tau only
s       0.0000    (12.0)       r    Delta only
r:s     0.0386    (12.0)       r    Delta only
i       0.0040    (4.0)        r    Delta only
cs      0.0147    (12.0)       r    Delta only
cr:s    0.0017    (12.0)       r    Delta only
ci      0.0004    (4.0)        r    Delta only
as:c    0.0232    (12.0)       dr   Delta and delta
ar:cs   0.0683    (12.0)       dr   Delta and delta
ai:c    0.0000    (4.0)        dr   Delta and delta
si      0.0000    (12.0*4.0)   r    Delta only
ri:s    0.0005    (12.0*4.0)   r    Delta only
csi     0.0002    (12.0*4.0)   r    Delta only
cri:s   0.0001    (12.0*4.0)   r    Delta only
asi:c   0.0002    (12.0*4.0)   dr   Delta and delta
ari:cs  0.0060    (12.0*4.0)   dr   Delta and delta

Results
s2(T)  =0.590
s2(D)  =0.264
s2(d)  =0.098
Er2    =0.858
Phi    =0.690

Computation sequence for D study
''a''   Differentiation   6.00
''c''   Stratification    3.00
''s''   Random           18.00
''r''   Random            1.00
''i''   Random            4.00
-------------------------------------------------------------
Pattern Var. Comp. Levels Signature Rule
-------------------------------------------------------------
c       0.1065    1            s    Delta only
a:c     0.5897    1            ds   tau only
s       0.0000    (18.0)       r    Delta only
r:s     0.0257    (18.0)       r    Delta only
i       0.0040    (4.0)        r    Delta only
cs      0.0098    (18.0)       r    Delta only
cr:s    0.0011    (18.0)       r    Delta only
ci      0.0004    (4.0)        r    Delta only
as:c    0.0155    (18.0)       dr   Delta and delta
ar:cs   0.0456    (18.0)       dr   Delta and delta
ai:c    0.0000    (4.0)        dr   Delta and delta
si      0.0000    (18.0*4.0)   r    Delta only
ri:s    0.0003    (18.0*4.0)   r    Delta only
csi     0.0001    (18.0*4.0)   r    Delta only
cri:s   0.0001    (18.0*4.0)   r    Delta only
asi:c   0.0002    (18.0*4.0)   dr   Delta and delta
ari:cs  0.0040    (18.0*4.0)   dr   Delta and delta
```

Appendix D

Error messages

As an aid in troubleshooting, we provide here a summary of all error messages of G_String IV. Each error message carries

a specific error code in {}. These identify uniquely, at which location of the code an error was detected.

Error of experimental design:

{E 10} Facets "Facet 1" and "Facet 2" are confounded. You would not get valid results!

Your experiment is poorly designed. You do not have a sufficient number of nested data in your study to resolve the confounding between it and the nested facet. G_String will deliver results, but they are meaningless.

Errors of design specification:

{D 10} Pattern should not be empty!

You have to define a design pattern for each nesting level. This error is fatal.

{D 20} G_String IV does not handle a subcomponent of type "x:y:z".

{D 21} G_String IV does not handle a subcomponent of type "x:y:z".

{D 22} At present, we does not handle effects of the type "x:y:z".

{D 24} At present, we does not handle effects of the type "x:y:z".

{D 25} G_String cannot handle this level of complexity at present.{x:y:z}.

These error messages all mean the same; they have been detected at various stages of calculation. G_String IV cannot handle this specific design complexity. Maybe, at a later stage we will figure out how to do it and will update the program. This error is fatal.

{D 30} You must have exactly one facet of differentiation!

{D 31} You must have exactly one facet of differentiation!

Under normal circumstances, you should not get this error, since following the steps of G_String will automatically prevent it. A corrupted, re-use control file, though, could give rise to this error. This error is fatal.

{D 40} Error in naming facets; typically duplication.

Each facet requires a distinct one-character abbreviation. This error is fatal.

{D 50} The facet of differentiation can only be nested in a facet of stratification.

Under normal circumstances, you should not get this error, since following the steps of G_String will automatically prevent it. A corrupted, re-use control file, though, could give rise to this error. This error is fatal.

Errors involving the control file:

{C 10} Control file is not well formed!

In order for G_String IV to re-use an existing control file, it has to be formed according to fixed rules (see p. 23 of the manual for an example). Specifically, the "comment" tag of the line specifying the facets must be terminate by a '%' character, i.e., "COMMENT%" rather than "COMMENT". When you use a control file generated by G_String_III or later, it is automatically in the correct format. This error is fatal.

Errors involving the data file:

(F 10} Data file "file name" is not readable.

The format of the file specified is not recognized as a data file format for either G_String or urGENOVA. This error is usually due to specifying the wrong file. This error is fatal.

{F 20} Data does not match facet specifications.

The facet specification doesn't correspond to the structure of the data file. Maybe, the asterisk was set to the wrong level (Step 7). This error is fatal.

{F 30} Insufficient records to calculate grand mean! Empty line "xxx."

{F 31} Data file does not contain sufficient data.

Either you require too many datapoints, or you dropped some data from your data file. This error is fatal.

{F 32} Your data file is missing "xxx" values. They have been replaced with the grand mean.

{F 33} Your data file is missing "xxx" values. They have been replaced with the grand mean.

These messages indicate that the structure of the data file is correct, but you have empty data cells. G_String will replace missing values with the grand mean, which is ok, if only a small percentage of cells are involved, and they are more or less randomly distributed through your data file. Otherwise you have to rethink your design, in order to avoid systematic errors.

{F 40} Unable to convert "String" to decimal number.

You may have mixed up your files, or left the column titles in the data file. G_String expects a numerical value, not characters. This error is fatal.

Internal errors:

{M 10} Crossed facets must have integer levels.

G_String expects that integer levels rather than fractional levels are specified for crossed facets. This error is fatal.

{M 20} Wrong averaging type "X!"

This error should not normally occur. G_String selects the appropriate averaging types according to rules listed in the manual and in Brennan. Theoretically, there could be internal errors that would call up an incorrect averaging type. This error is fatal.

Errors transmitted from urGENOVA:

{U 10} urGENOVA error: "message"

If urGENOVA fails for any reason, it emits an error message which is displayed by G_String. These errors are usually fatal.