Generalizability Theory for Clinician-Rated Outcomes

Joseph C. Cappelleri
Pfizer Inc, 445 Eastern Point Road (MS 8260-2502), Groton, CT 06340

Abstract

Generalizability (G) theory is a statistical theory about the dependability or reliability of behavioral or other measurements. The strength of G theory is that multiple sources of variance (such as from persons, raters, and time) in a measurement can be estimated separately in a single analysis. In the process, G theory provides a summary coefficient reflecting the level of dependability in measurement. Originating in the behavioral and educational sciences, G theory also has merit in the reliability assessment of clinical outcome assessments in the health sciences, in particular, for clinician-rated outcomes. Yet it has been underused there. With the advent of a published guidance on clinician-rated assessments of treatment benefit by an ISPOR Task Force, however, the application of G theory in this area is both timely and relevant. In this manuscript, the fundamentals and a pedagogical example of G theory in the context of clinician-rated outcomes are described and illustrated.

Key Words: generalizability theory, reliability, clinician-reported outcomes, clinical outcome assessments, measurement, dependability

1. Introduction

A clinical outcome assessment (COA) directly or indirectly measures how patients feel or function and can be used to determine whether a treatment has demonstrated efficacy, effectiveness, or safety (Cappelleri and Spielberg, 2015; Food and Drug Administration, 2017). A COA measures a specific concept (i.e., what is being measured by an assessment, such as pain intensity) within a particular context of use. There are four types of COAs: clinician-reported (clinician-rated) outcomes, patient-reported outcomes, observer-reported outcomes, and performance outcome assessments.

In this paper the focus is on clinician-rated outcomes which, like other COAs, can be influenced by human choices, judgement, or motivation. A clinician-rated assessment is conducted and reported by a trained health-care professional (Powers et al., 2017). Its proper assessment requires specialized professional training in order to evaluate the patient's health status. When a clinical interviewer injects his or her judgment in arriving at a score (e.g., regarding the patient's state of pain), then the type of COAs is clinician-reported outcome and not (say) a patient-reported outcome. An example would be a clinician-reported rating scale on the severity score of a patient's pain level over the past 24 hours using an 11-point numeric rating scale (e.g., 0 = no pain, 10 = worst possible pain).

Characterizing treatment benefit in terms that are meaningful and operationally sound is not only fundamental to clinical science, but also essential to the credibility and clarity in the communication of this vital information. In a report by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Task Force for Clinical Outcome Assessments, a clear conceptual foundation is provided for the development and evaluation of three types of clinician-reported outcome assessments: reading, rating, clinician global assessments (Powers et al., 2017).

Generalizability theory is a statistical theory about the dependability or reliability of behavioral or other measurements and has special appeal for measurement of clinician-rated outcomes. The strength of generalizability theory, which assumes that the measured attribute of interest (e.g., pain) is in a steady state, is that multiple sources of variance (such as from persons, raters, and time) in a measurement can be estimated separately in a single analysis. In the process, generalizability theory can provided a set of summary coefficients reflecting the level of dependability in measurement.

Generalizability theory was originally devised by Cronbach et al. (1972). The essence of the theory is the recognition that in any measurement situation there are multiple (in fact, infinite) sources of error variance. An important goal of measurement is to attempt to identify, measure, and thereby find strategies to reduce the influence of these sources on the measurement in question. The theory is about 45 years old and a handful of publications in the health sciences have applied generalizability theory (Streiner et al., 2015). Nevertheless, despite its potential, the application of generalizability theory remains limited and a relatively uncommon approach to measurement in the health sciences.

Originating in the behavioral and educational sciences, generalizability theory is especially suited for the reliability assessment of clinical outcome assessments in the health sciences – in particular for clinician-rated outcomes (Cappelleri et al. 2017; de Vet et al., 2011; Streiner et al., 2015). Yet it has been underused there. With the advent of a published guidance on clinician-rated assessments of treatment benefit by an ISPOR Task Force (Powers et al., 2017), however, the application of generalizability theory in this area is both timely and relevant. In this paper, the fundamentals and a pedagogical example of generalizability theory in the context of clinician-rated outcomes are described and illustrated.

2. Illustrative Description of Generalizability Theory

2.1 Some Basic Definitions

Generalizability (G) theory is a statistical theory for evaluating the dependability (or reliability) of behavioral and other measurements (Brennan, 2001; Shavelson and Webb, 1991). G theory permits the researcher to address a host of questions about the sources of measurement error and how it can be improved across a large variety of contexts. In this paper, dependability of measurement refers specifically to the accuracy of generalizing from a patient's observed score on the attribute of interest (e.g., depression) given by a clinician rater to the average score that patient would have received under all the possible conditions that the decision makers consider to be acceptable substitutes (e.g., scores of ratings sampled on Occasions 2 and 3) from the observation in hand (scores on Occasion 1 only).

Hence, in G theory, a rating score is conceived of as a sample from a *universe of admissible observations*, which a decision maker is willing to treat as interchangeable for the purposes of making a decision (such as, eventually, the impact of an intervention on clinician-rated depression). Implicit in this notion of dependability is the assumption that the patient's measured attribute is in a steady state. As such, it is assumed that differences among scores from a rater on the same individual on different occasions of measurement are due to one or more sources of error and not to systematic changes in what is being rated.

A *facet of* measurement is a characteristic feature such as rater, occasion (time), and items (for health scales with multiple items) in the design of the study. A universe of admissible observations, then, is defined by all possible combinations of the different levels of the facets (e.g., raters, occasions).

2.2. Pedagogical Example

Consider a generalizability study where physiotherapists rate the physical function on a group of respiratory patients on multiple occasions before the patients start rehabilitation treatment in a clinical trial. The intention is to assess and quantify the reliability of clinician-rated physical function, which is being considered as a measure in a subsequent clinical trial.

Suppose that patients are being initially considered and three therapists are involved in their rating of physical function (in a real-life application much more than five patients would be needed to obtain accurate results). Each therapist rates each patient's physical function on a 10-point scale (where higher scores signify better physical function) on two occasions at one and two weeks. In this G study, the patients are the *object of measurement* and both *raters* and *occasions* are *facets* of measurement.

The universe of admissible observations includes all possible raters and occasions that a decision maker would be equally willing to interpret as bearing on patients' physical function. Here, raters and occasion are considered *random facets*, so that therapists and occasions are assumed to be exchangeable with another sample of three therapists evaluated at another set of two occasions from the same admissible universe. Because each therapist rates each patient on the two occasions, these three factors are completely crossed in this design.

This two-facet design has seven sources of variability in a patient's rating (Table 1). One source of variability, attributable to the object of measurement, is individual differences among patients in their level of physical function. This source of variability is considered universe-score variance, the variability in the expected values of observed scores over all patients in the universe of generalization. The six other sources of variability are associated with the measurement facets and create inaccuracies in generalizing from the particular sample on measurement of physical function to the universe of admissible observations on the same patient.

Table 1. Sources of Variability in the Two-Facet Study on Physical Function

Source of Variability	*Type of Variation*	*Variance Notation*
Patient (*p*)	Universe-score variance (object of measurement)	σ_p^2
Raters (*r*)	Constant effect for all patients due to stringency of raters	σ_r^2
Occasions (*o*)	Constant effect for all patients due to inconsistencies from one occasion to another	σ_o^2
p x *r*	Inconsistencies of raters' evaluation on particular patients' physical function	σ_{pr}^2
p x *o*	Inconsistencies from one occasion to another in particular patients' physical function	σ_{po}^2
r x *o*	Constant effect for all patients due to differences in raters' stringency from one occasion to another	σ_{ro}^2
p x *r x o, e*	Residual consisting of the unique combination of *p*, *r*, *o*; plus unmeasured facets or random error that affect measurement, *e*	$\sigma_{pro,e}^2$

Inconsistencies among raters will create problems in generalizing from a patient's average score on physical function provided by a sample of three raters to his or her average score on physical function over all possible raters in the universe of admissible raters. Conclusions about a patient's physical function would depend on whether a rater tends to be liberal or stringent in her scoring. Note that the stringency of a rater applies to all patients in the population and, therefore, the rater effect is considered a main effect (in analysis of variance parlance), a constant effect uniform for all patients.

Similarly, inconsistencies in the level of physical function from one occasion to the next can also challenge generalization from sample to universe. Something that transpires on a particular occasion that affects all patients in the same way may increase or decrease their level of physical functioning. Hence this situation of a constant effect on all patients in the study would give rise to the occasion effect as a main effect.

In addition, inconsistencies in raters' scores of physical function may arise for particular patients. For example, relative to other raters, Rater 1 might be particularly liberal when rating subjects 1, 3 and 5, whereas Rater 2 might treat all subjects alike. Because of this variation, a person-by-rater interaction arises as only some patients and some raters in combination produce a distinctive result.

Likewise, some patients (but not all patients) may have higher levels of physical function on one occasion but not on another. This type of inconsistency, which is localized to particular patients and not all patients, gives rise to a patient-by-occasion interaction.

Another source of variability stems from the unique combination of rater and occasion interaction. For example, on one occasion, Rater 1 might be lenient in rating physical function for all patients, while on another occasion he might not be.

Finally, the last source of variability is the residual that includes the unique combination of patient, rater, and occasion (the patient-by-rater-by-occasion interaction) plus unmeasured sources of variation and random events or error.

G theory acknowledges that an assessment might be adapted for particular decisions and, in doing so, distinguishes a G study from a *decision (D) study*. In a G study, the universe of admissible observations is defined as broadly as possible, for example with respect to raters and occasions, in order to provide variance component estimates to a wide variety of decision makers.

A D study, on the other hand, typically selects only certain levels of the facets for a particular purpose, thereby narrowing the score interpretation to a more restricted universe of generalization. A different generalizability (reliability) coefficient can then be calculated for each specific purpose. In the physical function example, for instance, it might be decided to use three occasions (instead of two) and two raters (instead of three) for decision-making purposes; as such, the G coefficient could be calculated to reflect this proposed implementation.

3. Illustration of Generalizability Theory

A G study is designed specifically to isolate and estimate those facets of measurement error considered relevant for the research investigation. The study includes important facets that decision makers may wish to generalize over, such as raters and occasions.

Typically, “crossed” designed are used where all individuals are measured on all levels of all facets.

Let’s reconsider our two-facet illustrative (synthetic) example. Here five patients in the sample were rated on a 10-point scale (higher scores suggest higher levels of physical function) by three raters on two occasions at weeks 1 and 2 (Table 2). A crossed design provides maximum information about the variation contributed by the object of measurement (true-score or universe variance among patients), the facets, and their combinations to the total amount of variation in the observed scores.

Table 2. Data on Two-Facet Study on Physical Function

				Occasion		
	Week One			Week Two		
Rater	*R1*	*R2*	*R3*	*R1*	*R2*	*R3*
Patient						
1	4	5	4	2	3	3
2	2	1	3	3	4	3
3	0	1	0	2	2	3
4	5	4	4	4	4	4
5	3	3	3	1	3	2

As stated previously, the universe of generalization is defined as the set of facets and their levels (e.g., raters and occasions) to which a decision maker wants to generalize. A patient’s universe score (denoted as μ_p) is defined as the long-run average or, more technically, the expected value of his or her observed scores over all observations in the universe of generalization.

3.1 Model Development

In our two-facet crossed *p* x *r* x *o* (patient-by-rater-by-occasion) design, raters and occasions are considered random effects, because they are considered exchangeable with other raters and occasions from the universe of generalization. The object of measurement – patients – is not a source of error and, therefore, is not a facet. In the *p* x *r* x *o* design with generalization over all admissible raters and occasions taken from an indefinitely large universe, the components of an observed score (Y_{pro}) for a particular person (*p*) on a particular rater (*r*) and occasion (*o*) are as follows:

$$
\begin{array}{lll}
Y_{pro} = & & \\
 & \mu & \text{[grand mean]} \\
+ & \mu_p - \mu & \text{[patient effect]} \\
+ & \mu_r - \mu & \text{[rater effect]} \\
+ & \mu_o - \mu & \text{[occasion effect]} \\
+ & \mu_{pr} - \mu_p - \mu_r + \mu & \text{[patient-by-rater effect]} \\
+ & \mu_{po} - \mu_p - \mu_o + \mu & \text{[patient-by-occasion effect]} \\
+ & \mu_{ro} - \mu_r - \mu_o + \mu & \text{[rater-by-occasion effect]} \\
+ & Y_{pro} - \mu_{pr} - \mu_{po} - \mu_{ro} + \mu_p + \mu_r + \mu_o - \mu & \text{[residual effect],}
\end{array}
$$

where $\mu = E_p E_r E_o(Y_{pro})$ and $\mu_p = E_r E_o(Y_{pro})$ with E meaning expectation and other terms in the equation defined analogously.

Under the assumption of a random-effects model, the distribution of each component or “effect,” except for the grand mean, has a mean of zero and a variance component. The variance component for the person effect is $\sigma_p^2 = E_p(\mu_p - \mu)^2$, the universe-score variance. The variance components for the other effects are defined similarly. The residual variance component, $\sigma_{pro,e}^2$, reflects the person-rater-occasion interaction confounded with random error because there is only one observation per cell. The collection of observed scores, Y_{pro}, has a total variance that equals the sum of the variance components:

$$\sigma_{pro}^2 = \sigma_p^2 + \sigma_r^2 + \sigma_o^2 + \sigma_{pr}^2 + \sigma_{po}^2 + \sigma_{po}^2 + \sigma_{pro,e}^2.$$

Each variance component can be estimated from an analysis of variance framework using one of several options (e.g., least squares, maximum likelihood, restricted maximum likelihood, minimum variance quadratic variance). For the example in Table 2, a three-way random-effects analysis of variance model on person, rater, and occasion was performed in SAS using least squares estimation to estimate the variance components from the mean squares (Table 3). The relative magnitudes of the estimated variance components, except for σ_p^2, provide information about potential sources of measurement error influencing the measurement on physical function from a rater on an occasion. Statistical tests are not used in G theory; instead, standard errors for variance component estimates can provide information about sampling variability of the estimated variance components.

Table 3. SAS Program in the Two-Facet Study on Physical Function

```
Proc Varcomp Method=Type1; /*This method gives least squares estimation*/
                    Class Pat Rater Occasion;
                    Model Score = Pat Rater Occasion
                    Pat*Rater  Pat*Occasion   Rater*Occasion;
Run;
```

3.2. Model Results

In our example, the estimated patient (universe-score) variance (27.2%) was substantial and indicates that, when averaged over raters and occasions, patients in the sample differed systematically in their physical function (Table 4). Hence, because patients constitute the object of measurement, not measurement error, this variability represents systematic individual differences in physical function. In addition, there was much patient-by-occasion interaction (47.5%), indicating that the relative standing of patients differed by occasion (time). Thus, a patient who showed scored high on physical function at one time did not necessarily score high at another occasion. This result implies that more time points are needed.

Table 4. Sources of Estimated Variability in the Two-Facet Study on Physical Function

Source	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*Variance Component*	*Percent of Total Variability*
Patient (*p*)	27.67	4	6.92	0.59	27.2
Rater (*r*)	0.87	2	0.43	0.03	1.4
Occasion (*o*)	0.03	1	0.03	0.00	0.0
p x *r*	2.13	8	0.27	0.00	0.0
p x *o*	14.47	4	3.62	1.03	47.5
r x *o*	0.87	2	0.43	0.00	0.0
Residual	4.13	8	0.52	0.52	23.9
Total	50.17	29		2.17	100.0

Note: The estimated variance components for *o, p* x *r,* and *r* x *o* were negative and set to zero.

Other interactions related to *p* x *r* and *r* x *o* accounted for no variation. There was also no variation between occasions, indicating that physical function was stable across occasions when physical function scores were averaged across patients and raters. Variability in raters accounted for little variation in scores (1.4%). On the other hand, the residual variance was relatively high (23.9%), which is reflective of the varying relative standing of patients across raters and occasions or other sources of errors (or a combination thereof) not systematically incorporated into the G study.

3.3 Generalizability Coefficients

The results of our illustrative G study can be used to optimize the number of levels (or conditions) of each facet in order to obtain a desired level of reliability (generalizability). The optimal number of conditions of a facet may be less than, equal to, or greater than the number of the conditions in the G study. The subsequent D study can be targeted with the optimal number of such levels, along with the appropriate study design, in order to arrive at the desired interpretation in the D study (relative vs. absolute interpretations).

G theory distinguishes decisions based on the relative standing or ranking of individuals (*relative* interpretations) and decisions based on the absolute level of their scores (*absolute* interpretations). For instance, the decision maker may want to know the correlation between physical function scores and, say, mental function scores (a relative interpretation). Or interest may lie in assigning all patients who have eventually attained a certain level of mastery to an advanced form of physiotherapy (an absolute decision).

These different interpretations of measurement affect the definitions of error and generalizability (reliability) coefficients, which range from 0 to 1 (with higher values indicative of higher reliability). For *relative decisions,* all variance components that influence the relative standing of individuals contribute to error (e.g., how well patients compare to each other in their physical function); for *absolute decisions*, all variance components except the object of measurement contribute to measurement error.

For our two-facet study, where patients were crossed with raters and occasions, (*p* x *r* x *o*), the formula for the relative reliability coefficient is given by

$$\frac{\sigma_p^2}{[\sigma_p^2 + (\sigma_{po}^2 / o') + (\sigma_{pr}^2 / r') + (\sigma_{res}^2 /(o'*r'))]}$$

.

Therefore, the dependability of the relative scores on physical function for a single measurement with one occasion (*o’*=1) and one rater (*r’*=1) equals 0.28 = (0.59) / (0.59 + 1.03 + 0 + 0.52). For five occasions (*o’*=5) and two raters (*r*’=2), the relative coefficient increases to 0.70 = (0.59 / 0.85) when the physical function scores are averaged over five occasions and two raters.

For our two-facet study, the formula for the absolute reliability coefficient is given by

$$\frac{\sigma_p^2}{[\sigma_p^2 + (\sigma_r^2 / r') + (\sigma_o^2 / o') + (\sigma_{po}^2 / o') + (\sigma_{pr}^2 / o') + (\sigma_{ro}^2 / r' * \mathrm{o}') + (\sigma_{res}^2 /(o' * \mathrm{r}'))]}$$

As such, dependability of the absolute scores on physical function for a single measurement with one occasion (*o’*=1) and one rater (*r’*=1) equals 0.27 = (0.59 / 2.17). For five occasions (*o’*=5) and two raters (*r’*=2), the absolute coefficient increases to 0.68 = (0.59 / 0.86) when the physical function scores are averaged over five occasions and two raters.

4. Conclusions

In this paper, a purely pedagogical illustration was used throughout based on three therapist ratings of physical function on two different occasions for a group of only five respiratory rehabilitation patients. Variance components used in estimating reliability coefficients can be unstable depending on the number of patients studied (Cronbach et al. 1972). In a real-life application, however, many more than five patients would be needed to obtain stable results. For example, one investigation concluded that at least 50 individuals would be needed for unbiased estimation (Atilgan, 2013).

Multi-colored applications and interpretations abound in the use of generalizability theory for clinician-rated outcomes. Consider, for instance, a non-interventional methodological study designed to assess and quantify the reliability of a clinician-rated measure (say, cognitive function) considered in a clinical trial. Consider again multiple clinical raters scoring the same set of patients at multiple time points (a crossed two-facet study with rater and time as random effects).

Here, a generalizability study can address several elements on the reliability of measurement for the clinician-rated outcome by involving a comparison on the measurements of all patients performed by different clinicians, across clinicians but not across time (i.e., inter-rater reliability); a comparison on one measurement by one clinician with another measurement by another clinician, across clinicians and time; a comparison of measurements performed over time by the same clinician, across time measurements but not across clinicians (i.e., intra-rater reliability); and whether higher reliability is obtained from using the average of multiple measurements of a patient by one clinician or using the average of one measurement by different clinicians – all of which can be used to make a decision in the planning of a subsequent (decision) study.

These metrics of reliability can be obtained by using the appropriate formulas for the ratio of variance components (de Vet et al., 2011).

This paper is not intended to provide a detailed or comprehensive exposition on generalizability theory. Topics not covered here – such as different types of generalizability studies (beyond the crossed design with two random facets discussed here), different types of facets (such as items on a scale intended to measure the same attribute), random vs. fixed facets, sample size considerations, and different approaches for estimation of variance components – are discussed elsewhere (Brennan, 2001; Shavelson and Webb, 1991).

The intent of this paper, instead, is to motivate the topic of generalizability theory so that medical researchers involved with clinician-rated outcomes are made aware (or more aware) of its benefits. In doing so, these researchers would be in a better position to more regularly and effectively apply the methodology to improve the reliability of measurement on clinician-rated outcomes.

References

1. Atilgan, H. (2013). Sample Size for Estimation of G and Phi Coefficients in Generalizability Theory. *Eurasian Journal of Educational Research*, 51, 215-228.

2. Brennan, R.L. (2001). *Generalizability Theory*. New York, NY: Springer-Verlag.

3. Cappelleri, J.C, Spielberg, S.P. (2015). Advances in Clinical Outcome Assessments. *Therapeutic Innovation & Regulatory Science*, 49, 780-782

4. Cappelleri, J.C., Deal, L.S., Petrie, C.D. (2017). Editorial. Reflections on ISPOR’s Clinician-Reported Outcomes Good Measurement Practice Recommendations. *Value in Health*, 20, 15-17.

5. Cronbach, L.J., Gleser, G. C., Nanda, H., Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability of Scores and Profiles*. New York, NY: John Wiley & Sons.

6. de Vet, H.C.W., Terwee, C.B., Mokkink, L.B., Knol, D.L. (2011). *Measurement in Medicine: A Practical Guide*. New York, NY: Cambridge University Press.

7. Food and Drug Administration. *Clinical Outcome Assessment Qualification Program*. Accessed online, September 9, 2017. https://www.fda.gov/drugs/developmentapprovalprocess/drugdevelopmenttoolsqualificationprogram/ucm284077.htm.

8. Powers, J.H. III, Patrick, D.L., Walton, M.K., Marquis, P., Cano, S., Hobart, J., Isaac, M. Vamvakas, S., Slagle, A., Molsen, E., Burke, L.B. (2017). Clinician-Reported Outcome Assessments of Treatment Benefit: Report of the ISPOR Clinical Outcome Assessment Emerging Good Practices Task Force. *Value in Health,* 20, 2-14.

9. Steiner, D.L., Norman, G.R., Cairney, J. (2015). *Health Measurement Scales: A Practical Guide to Their Development and Use*. Fifth edition. New York, NY: Oxford University Press.

10. Shavelson, R.J., Webb, N.M. (1991). *Generalizability Theory: A Primer*. Newbury Park, California: SAGE Publications.