# EPSY 5221 Generalizability Theory

G-Theory makes no assumptions regarding parallel measurements. In Classical Test Theory, we can only estimate one source of error at a time. G-Theory estimates the magnitude of each source of error and provides a mechanism for optimizing the reliability of other measurement. G-Theory also provides two reliability indexes – Generalizability coefficient (classical relative-reliability) and Dependability coefficient (absolute reliability).

*Framework-Definitions*

 Object of Measurement: the factor we are primarily focusing on, usually persons.

 Universe of Generalization: the whole collection of possible observations to which we wish to generalize.

 Universe Score: equivalent to the True-Score, it is the average of measurements in a universe to which we generalize.

 Facets: a set of measurement conditions; aspects of the measurement procedure; a potential source of error in generalization.

 G-Study: the procedures used to collect information on as many aspects of a measurement procedure as possible, to estimate variance components for one unit of each facet.

 D-Study: makes use of the information provided by the G-Study to design the best measurement procedure – minimizing undesirable source of error and maximizing reliability.

*Fixed/Random*

 Fixed facets contain only the conditions we are interested in using.

 Random facets contain a set of conditions randomly sampled from the universe of conditions for that facet, considered exchangeable with any other sample.

*Crossed/Nested*

 Crossed designs are those where every conditions of each facet is repeated for every condition of the other facets (e.g., every rater rates every item)

 Nested designs are those where certain conditions of a facet are only associated with certain conditions of another facet (e.g., each item is rated by a different group of raters—raters are nested in items)

*Generalizability/Dependability Coefficients*

***G*-Coefficient**

 A measure of generalizability for relative decisions: , which is approximately equal to the expected value of the squared correlation between observed scores and universe scores. It is used for relative reliability where we are only concerned with ranking individuals and not with task difficulty; the coefficient uses all sources of error that influence the relative standing of individuals—the interactions of each facet with the object of measurement.

***D*-Coefficient**

 A measure of dependability for absolute decisions: . It is used for absolute reliability (especially important when making criterion-referenced decisions) and includes all variance components except the object of measurement.

*G-Study Design Notes*

1. specify as many facets as possible based on major sources of error most likely present in a measurement procedure
2. use as many crossed facets as possible – avoid the nested design
3. if fixed facets are used and the conditions of those facets are not exchangeable, we should do separate G-studies for each condition

*Limitations*

1. assumptions are hard to achieve, including defining the universe and randomly sampling from that universe
2. sample dependent; dealing with observed scores
3. doesn’t provide information regarding specific conditions of a given facet (items, persons)
4. assumes the conditions are exchangeable—ignores maturation within a facet
5. defining universe of generalization
6. technical problems: dealing with missing data, dealing with ordered facets

**An Example**

Consider an example for a design that is completely crossed: 30 students are all rated by three raters on four tasks. An analysis of variance summary table for our example design could be as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | *ss* | *df* | *ms* | variance component |
|  |  |  |  |  |
| Persons (p) | 79.52 | 29 | 2.742 | 0.140 |
| Raters (r) | 2.84 | 2 | 1.420 | 0.005 |
| Tasks (t) | 6.96 | 3 | 2.319 | 0.009 |
| Persons x Raters (pr) | 20.18 | 58 | 0.348 | 0.022 |
| Persons x Tasks (pt) | 84.87 | 87 | 0.975 | 0.239 |
| Raters x Tasks (rt) | 4.53 | 6 | 0.754 | 0.017 |
| Persons x Raters x Tasks (prt) | 45.14 | 174 | 0.259 | 0.259 |

|  |  |  |
| --- | --- | --- |
| Effect  | Formula | Example |
|  |  |  |
| p | (MSp – MSpr – MSpt + MSprt) / NrNt | (2.742 - .348 - .975 + .259) / 12 = 0.1400 |
| r | (MSr – MSrt – MSpr + MSprt) / NpNt | (1.420 - .754 - .348 + .259) / 120 = 0.0050 |
| t | (MSt – MSrt – MSpt + MSprt) / NpNr | (2.319 - .754 - .975 + .259) / 90 = 0.0090 |
| pr | (MSpr – MSprt) / Nt | (.348 - .259) / 4 = 0.0220 |
| pt | (MSpt – MSprt) / Nr | (.975 - .259) / 3 = 0.2390 |
| rt | (MSrt – MSprt) / Np | (.754 - .259) / 30 = 0.0165 |
| prt | MSprt | 0.2590 |

In the following models, we are interested in differentiation of individuals. First we assume the levels of all three facets are randomly selected from some population. That is, the three raters in our study were selected from a large pool or universe of available raters as were subjects and tasks.

If we want to generalize to similar situations in which three randomly selected raters and four randomly selected items were used to supply ratings on a group of people, the formula, computation and generalizability coefficient would be as follows:



For decisions based on one rater and four tasks: 

For decisions based on three raters and one task: 

For decisions based on a single task and rater: 

Now let’s assume that we want to generalize to a new situation with the same set of raters or tasks; that is, raters and/or tasks are considered fixed, not randomly selected. First consider tasks fixed:



When raters are considered fixed:



When both are fixed:



Various conceptualizations of generalizability can be visualized in terms of Venn diagrams.

NRNT VCP

(3 x 4 x 0.14)

NPNT VCR

(30 x 4 x 0.005)

NT VCPR

(4 x 0.022)

NR VCPT

(3 x 0.239)

NP VCRT

(30 x 0.017)

VCPRT

(0.259)

NPNR VCT

(30 x 3 x 0.009)

Mean Square TASKS

Mean Square PERSONS

Mean Square RATERS

When reliability associated with the random model and generalizability to another sample of four items and three raters is the relevant question, the numerator is the portion of the person circle which doesn’t overlap with any other circle. The denominator is composed of the entire person circle.

From R. L. Brennan, *Generalizability Theory*, An NCME Instructional Module

Suppose an investigator, Mary Smith, decides that she wants to construct one or more measurement procedures for evaluating writing proficiency. She might proceed as follows.

She might identify or otherwise characterize, essay prompts that she would consider using, as well as potential raters of writing proficiency. She is not committing herself to actually using any specific items or raters. She is merely characterizing the facets of measurement that might interest her or other investigators. A facet is simply a set of similar conditions of measurement. Specifically, Smith is saying that any one of the essay prompts constitutes an admissible (i.e., acceptable to her) condition of measurement for her essay-prompt facet. Similarly, only one of the raters constitutes an admissible condition of measurement for her rater facet. We say that Smith’s universe of admissible observations contains an essay-prompt facet and a rater facet.

Further, suppose Smith would accept as meaningful to her a pairing of any rater (r) with any prompt (t). If so, Smith’s universe of admissible observations would be described as being crossed, and it would be denoted t x r, where the x is read “crossed with.” Specifically, if there were Nt prompts and Nr raters in Smith’s universe, then it would be described as crossed if any one of the NtNr combinations of conditions from the two facets would be admissible for Smith. Here it will be assumed that Nt and Nr are both very large – approaching infinity, at least theoretically.

G-theory does not presume that there is some particular definition of prompt and rater facets that all investigators would accept. Smith may characterize the potential raters as college instructors with a PhD in English – others may consider a rater facet consisting of high school teachers of English.

The word *universe* is reserved for conditions of measurement (prompts and raters) whereas the word *population* is reserved for the objects of measurement (persons).

Smith accepts as admissible the response of any person in the population to any prompt in the universe evaluated by any rater in the universe. If so, the population and universe of admissible observations are crossed, which is represented by p x (t x r) or simply p x t x r.

Using the notation of analysis of variance, any observable score for a single essay prompt evaluated by a single rater can be represented as:

*Xptr* = μ + α*p* + α*t* + α*r* + α*pt* + α*pr* + α*tr* + α*ptr*

P

R

PR

PT

PTR

RT

T

## Venn diagram for a p x t x r design

The variances of the scores, over the population of persons and the conditions in the universe of admissible observations is:

σ2(Xptr) = σ2(p) + σ2(t) + σ2(r) + σ2(pt) + σ2(pr) + σ2(tr) + σ2(ptr)

The total observed score variance can be decomposed into seven independent variance components. It is assumed here that the population and both facets in the universe of admissible observations are infinite. Under these assumptions, the variance components are called random effects variance components. The variance components are for single person-prompt-rater combinations, as opposed to average scores over prompts and/or raters.

Now that Smith has specified her population and universe of admissible observations, she needs to collect and analyze data to estimate the variance components. She obtains a sample of nr raters who use a particular scoring procedure to evaluate each of the responses by a sample of np persons to a sample of nt essay prompts. This study is called a G (Generalizability) Study.

What results are estimates of actual variances (parameters) from the equation above. We can interpret these in the following way:

Suppose that, for each person in the population, we could obtain each person’s mean score (technically, expected score) over all Nt essay prompts and all Nr raters in the universe of admissible observations. The variance of these mean scores (over the population of persons) is σ2(p). Each variance component can be interpreted in this way.

Interaction variance components are more difficult to interpret verbally, but approximate statements can be made. For example, σ2(pt) estimates the extent to which the relative ordering of persons differs by essay prompt, and σ2(pr) estimates the extent to which persons are rank ordered differently by different raters.

Notice in our example, the variance attributable to tasks (0.009) is nearly twice that attributable to raters (0.005). This suggests that prompts differ much more in average difficulty than raters differ in average stringency.

Considering the interaction components, the person by task term (0.239) is ten times as large as the person by rater term (0.022). In conjunction with the earlier result, we can see that tasks are a considerably greater source of variability in persons’ scores than are raters.

# D Study Considerations

The estimates we obtain from our G studies provide information we can use to design efficient measurement procedures for operational use – for making substantive decisions about objects of measurements in various D (decision) studies.

Suppose that Smith decides to design her measurement procedure such that each person will respond to nt essay prompts with each response to every prompt evaluated by the same nr raters. Furthermore, the decisions about a person will be based on his or her mean score of the ntnr observations associated with the person. This is the p x T x R design for a D study.

The sample sizes do not need to be the same in the D study as in the G study. The D study focuses on mean scores for persons, rather than single person-prompt-rater observations focused on in the G study; thus the use of upper case letters.

The *universe of generalization* can be conceptualized as a universe of measurement procedures each employing the specified D study sample sizes and design structure. In G theory these measurement procedures are described as randomly parallel, and it is assumed that any particular measurement procedure consists of a random sample of conditions for at least one facet.

The *universe score* is the expected value of the mean scores for a person from every instance of the measurement procedure in the universe of generalization. The variance of universe scores over all persons in the population is called universe score variance. This is conceptually similar to true score variance in CTT.

D Study variance components can be obtained from the G study components. For example, consider the following G study variance components:

|  |  |  |
| --- | --- | --- |
| *Effect*  | *G-Study Variance Component* | *D-Study for 3 tasks and 2 raters* |
|  |  |  |
| p | 0.25 | 0.25 |
| r | 0.02 | 0.01 |
| t | 0.06 | 0.02 |
| pr | 0.04 | 0.02 |
| pt | 0.15 | 0.05 |
| rt | 0.00 | 0.00 |
| prt | 0.12 | 0.02 |

RULE: given a G-study estimated variance component, to obtain the D-study variance components, divide the G-study component by nt if it contains ***t*** but not ***r***, by nr if it contains ***r*** but not ***t***, and by ntnr if it contains both ***t*** and ***r***.

A person’s universe score is defined as μp = μ + αp and the variance of universe scores is σ2(τ), in our scenario, σ2(p).

# Error Variances

*Absolute error variance*, σ2(Δ), is simply the difference between a person’s observed and universe scores: Δp = XpTR - μp

Given our scenario above, Δp = α*T* + α*R* + α*pT* + α*pR* + α*TR* + α*pTR*

The variance of the absolute errors, σ2(Δ), is the sum of all the variance components except the variance for p.

In our example, σ2(Δ) = 0.02 + 0.01 + 0.05 + 0.02 + 0.00 + 0.02 = 0.12

And its square root is 0.35; this is interpretable as an estimate of the absolute standard error of measurement (can be used to create a confidence interval for the person’s universe score).

*Relative error variance*, σ2(δ), is defined as the difference between a person’s observed deviation score and his or her universe deviation score: δp = (XpTR - μTR) – (μp - μ).

For our scenario, δp = α*pT* + α*pR* + α*pTR*

Relative error variance is similar to error variance in CTT. For our case of three tasks and 2 raters, σ2(δ) = 0.05 + 0.02 + 0.02 = 0.09. Its square root is 0.30, which is interpretable as an estimate of the relative standard error of measurement.

In general, relative error variance is less than absolute error variance because it includes fewer variance components. This suggests that relative interpretations about persons’ scores are less error prone than absolute interpretations.

# Coefficients

Two types of reliability-like coefficients are available in generalizability theory.

*Generalizability Coefficient*, ρ2.

A generalizability coefficient is defined as rho-squared: .

In our scenario, = .74

*Index of Dependability*, Φ.

An index of dependability is defined as phi: .

In our scenario, = .68

Φ involves absolute error variance whereas ρ2 involves relative error variance. Consequently, Φ is generally less than ρ2. The index, Φ, is appropriate when scores are given absolute interpretations, as in domain-referenced or criterion-reference situations. The index ρ2 is appropriate when scores are given relative interpretations, as in norm-referenced situations.

Other forms of reliability can be estimated as well:

Inter-Rater Reliability 

Internal Consistency 