5

Generalizability Theory

Up to this point, we have discussed reliability from the perspective of classical test theory (CTT). However, as we discussed, CTT makes several assumptions that are unrealistic. For example, CTT assumes that all error is random. In CTT, there is an assumption that there exists a true score that is an accurate measure of the trait under a *specific* set of conditions. There are other measurement theories that conceptualize reliability differently than the way that CTT conceptualizes reliability. One such measurement theory is generalizability theory (Brennan, 1992), also known as G-theory and domain sampling theory. G-theory is also discussed in the chapters on reliability (Chapter 3, Section 3.11), validity (Chapter 4, Section 4.7) and structural equation modeling (Chapter 7, Section 7.11).

5.1 Overview

G-theory is an alternative measurement theory to CTT that does not treat all measurement differences across time, rater, or situation as "error" but rather as a phenomenon of interest (Wiggins, 1973). G-theory is a measurement theory that is used to examine the extent to which scores are consistent across a specific set of conditions. In G-theory, the true score is conceived of as a person's *universe score*—the mean of all observations for a person over all conditions in the universe—this allows us to estimate and recognize the magnitude of multiple influences on test performance. These multiple influences on test performance are called *facets*.

5.1.1 The Universe of Generalizability

Instead of conceiving of all variability in a person's scores as error, G-theory argues that we should describe the details of the particular test situation (universe) that lead to a specific test score. The universe is described in terms of its facets:

- settings
- observers (e.g., amount of training they had)
- instruments (e.g., number of items in test)
- occasions (time points)
- attributes (i.e., what we are assessing; the purpose of test administration)

Measures with strong reliability show a high ratio of variance as a function of the person relative to the variance as a function of other facets or factors. To the extent that variance in scores is attributable to different settings, observers, instruments, occasions, attributes, or other facets, the reliability of the measure is weakened.

TABLE 5.1 Percent of Variance from Different Sources in Generalizability Theory Model with Three Facets: Person, Item, and Occasion (and Their Interactions). (Adapted from Webb et al. (2005), table 1, p. 2. Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: Overview. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 717–719). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013192.bsa703)

Source	Variance Accounted For (%)
Person (p)	30
Item (i)	5
Occasion (o)	3
p x i	25
p x o	5
i x o	2
p x i x o	10
residual	20

5.1.2 Universe Score

A person's universe score is the average of a person's scores across all conditions in the universe. According to G-theory, given the exact same conditions of all the facets in the universe, the exact same test score should be obtained. This is the universe score, which is analogous to the true score in CTT.

5.1.3 G-Theory Perspective on Reliability

G-theory asserts that the reliability of a test does not reside within the test itself; a test's reliability depends on the circumstances under which it is developed, administered, and interpreted. A person's test scores vary from testing to testing because of (many) variables in the testing situation. By assessing a person in multiple facets of the universe, this allows us to estimate and recognize the magnitude of multiple sources of measurement error. Such measurement error includes:

- day-to-day variation in performance (stability of the construct, test–retest reliability)
- variance in the item sampling (coefficient of internal consistency)
- variance due to both day-to-day and item sampling (coefficient of equivalence from parallel-forms reliability, or convergent validity, as discussed in Section 4.7)

In G-theory, all sources of measurement error (facets) are considered simultaneously—something CTT cannot achieve (Shavelson et al., 1989). This occurs through specifying many different variance facets in the estimation of the true score, rather than just one source of error variance as in CTT. This specification allows us to take into consideration variance due to occasion effects, item effects, and occasion × item effects (i.e., main effects of the facets in addition to their interaction), as in Table 5.1.

A score's usefulness largely depends on the extent to which it allows us to generalize accurately to behavior in a wider set of situations—i.e., a universe of generalization. The G-Theory equivalent of the CTT reliability coefficient of a measure is the generalizability coefficient or dependability coefficient.

5.1.4 G-Theory Perspective on Validity

G-theory can simultaneously consider multiple aspects of reliability and validity in the same model. For instance, internal consistency reliability, test–retest reliability, inter-rater reliability, parallel-forms reliability, and convergent validity (in the D. T. Campbell & Fiske, 1959 sense of the same construct assessed by a different method) can all be incorporated into a G-theory model.

For example, a G-theory model could assess each participant across the following facets:

- time: e.g., T1 and T2 (test–retest reliability)
- items: e.g., questions within the same instrument (internal consistency reliability) and questions across different instruments (parallel-forms reliability)
- rater: e.g., self-report and other-report (inter-rater reliability)
- method: e.g., questionnaire and observation (convergent validity)

Using such a G-theory model, we can determine the extent to which scores on a measure generalize to other conditions, measures, etc. Measures with strong convergent validity show a high ratio of variance as a function of the person relative to the variance as a function of measurement method. To the extent that variance in scores is attributable to different measurement methods, convergent validity is weakened.

An example data structure that could leverage G-theory to partition the variance in scores as a function of different facets (person, time, item, rater, method) and their interactions is in Table 5.2.

TABLE 5.2 Example Data Structure for Generalizability Theory with the Following Facets: Person, Time, Item, Rater, Method.

Person	Time	Item	Rater	Method	Score
1	1	“hits others”	1	questionnaire	10
1	1	“hits others”	1	observation	15
1	1	“hits others”	2	questionnaire	8
1	1	“hits others”	2	observation	13
1	1	“argues”	1	questionnaire	4
1	1	“argues”	1	observation	2
1	1	“argues”	2	questionnaire	5
1	1	“argues”	2	observation	7
1	2	“hits others”	1	questionnaire	8
1	2	“hits others”	1	observation	10
1	2	“hits others”	2	questionnaire	6
1	2	“hits others”	2	observation	7
1	2	“argues”	1	questionnaire	2
1	2	“argues”	1	observation	2
1	2	“argues”	2	questionnaire	4
1	2	“argues”	2	observation	6
2	1	“hits others”	1	questionnaire	5
...	...	...	...	...	...

In sum, G-theory can be a useful way of estimating the degree of reliability and validity of a measure’s scores in the same model.

5.1.5 Generalizability Study

In G-theory, the goal is to conduct a generalizability study (G study) and a *decision study* (D study). A generalizability study examines the extent of variance in the scores that is attributable to various facets. The researcher must specify and define the universe (set of conditions) to which they would like to generalize their observations and in which they would like to study the reliability of the measure. For instance, it might involve randomly sampling from within that universe in terms of people, items, observers, conditions, timepoints, measurement methods, etc.

5.1.6 Decision Study

After conducting a generalizability study, one can then use the estimates of the extent of variance in scores that are attributable to various facets (estimated from the generalizability study) to conduct a decision study (D study). A decision study examines how generalizable scores from a particular test are if the test is administered in different situations. In G-theory, reliability is estimated with the *generalizability coefficient* and the *dependability coefficient*.

5.1.7 Analysis Approach

Traditionally, a generalizability theory approach would test the generalizability study and decision study using a factorial analysis of variance [ANOVA; Brennan (1992)], as exemplified in Section 5.2.4.1, (as opposed to simple ANOVA in CTT). However, ANOVA is limiting—it works best with balanced designs, such as with the same sample size in each condition/facet; but in most real-world applications, data are not equally balanced in each condition. So, it is better to fit G-theory models in a mixed model framework, as exemplified in Section 5.2.4.2.

5.1.8 Practical Challenges

G-theory is strong theoretically, but it has not been widely implemented. G-theory can be challenging because the researcher must specify, define, and assess the universe to which they would like to generalize their observations and to understand the reliability of the measure.

5.2 Getting Started

5.2.1 Load Libraries

```r
library("petersenlab")
library("gtheory")
library("MOTE")
library("here")
library("tidyverse")
library("tinytex")
library("knitr")
library("kableExtra")
library("rmarkdown")
library("bookdown")
```

5.2.2 Prepare Data

5.2.2.1 Generate Data

```
set.seed(52242)

Person <- as.factor(rep(1:6, each = 8))
Occasion <- Rater <- as.factor(rep(1:2, each = 4, times = 6))
Item <- as.factor(rep(1:4, times = 12))
Score <- c(
  9,9,7,4,9,8,5,5,9,8,4,6,
  6,5,3,3,8,8,6,2,8,7,3,2,
  9,8,6,3,9,6,6,2,10,9,8,7,
  8,8,9,7,6,4,5,1,3,2,3,2)
```

5.2.2.2 Add Missing Data

Adding missing data to dataframes helps make examples more realistic to real-life data and helps you get in the habit of programming to account for missing data.

```
Score[30] <- NA
```

5.2.2.3 Combine Data into Dataframe

```
pio_cross_dat <- data.frame(Person, Item, Score, Occasion)
```

Below are examples implementing G-theory. The `pio_cross_dat` data file for these examples comes from the `gtheory` package (Moore, 2016). The examples are adapted from Huebner & Lucht (2019).

5.2.3 Universe Score for Each Person

Universe scores for each person are generated using the following syntax and are presented in Table 5.3.

```
universeScores <- pio_cross_dat %>%
  group_by(Person) %>%
  summarise(universeScore = mean(Score, na.rm = TRUE), .groups = "drop")
```

5.2.4 Generalizability (G) Study

Generalizability studies can be conducted in an ANOVA or mixed model framework. Below, we fit a generalizability study model in each framework. In these models, the item, person, and their interaction appear to be the three facets that account for the most variance in scores. Thus, when designing future studies, it would be important to assess and evaluate these facets.

TABLE 5.3 Participants' Universe Scores.

Person	Universe Score
1	7.000
2	5.500
3	5.500
4	6.143
5	8.250
6	3.250

5.2.4.1 ANOVA Framework

```
summary(aov(
  Score ~ Person*Item*Occasion,
  data = pio_cross_dat))
```

```
                     Df Sum Sq Mean Sq
Person                5  113.0    22.6
Item                  3  119.2    39.7
Occasion              1   14.2    14.2
Person:Item          15   35.6     2.4
Person:Occasion       5    6.6     1.3
Item:Occasion         3    2.3     0.8
Person:Item:Occasion 14   12.1     0.9
1 observation deleted due to missingness
```

5.2.4.2 Mixed Model Framework

The mixed model framework for estimating generalizability is described by Jiang (2018).

```
summary(lmer(
  Score ~ 1 + (1|Person) + (1|Item) + (1|Occasion) +
    (1|Person:Occasion) + (1|Person:Item) + (1|Occasion:Item),
  data = pio_cross_dat))
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Score ~ 1 + (1 | Person) + (1 | Item) + (1 | Occasion) + (1 |
    Person:Occasion) + (1 | Person:Item) + (1 | Occasion:Item)
   Data: pio_cross_dat

REML criterion at convergence: 174.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.3023 -0.6530 -0.0525  0.6356  1.3702

Random effects:
 Groups          Name        Variance Std.Dev.
 Person:Item     (Intercept) 0.78731  0.8873
 Person:Occasion (Intercept) 0.11728  0.3425
 Occasion:Item   (Intercept) 0.00157  0.0397
 Person          (Intercept) 2.47804  1.5742
 Item            (Intercept) 3.12380  1.7674
 Occasion        (Intercept) 0.53663  0.7326
 Residual                    0.83743  0.9151
Number of obs: 47, groups:
Person:Item, 24; Person:Occasion, 12; Occasion:Item, 8; Person, 6; Item, 4; Occasion, 2

Fixed effects:
            Estimate Std. Error t value
(Intercept)     5.96       1.23    4.83
```

```
PxIxO <- gstudy(
  data = pio_cross_dat,
  Score ~ (1|Person) + (1|Item) + (1|Occasion) + (1|Person:Item) +
    (1|Person:Occasion) + (1|Occasion:Item))

PxIxO
```

```
$components
           source      var percent n
1     Person:Item 0.787305    10.0 1
2 Person:Occasion 0.117284     1.5 1
3   Occasion:Item 0.001574     0.0 1
4          Person 2.478043    31.4 1
5            Item 3.123803    39.6 1
6        Occasion 0.536633     6.8 1
7        Residual 0.837426    10.6 1

attr(,"class")
[1] "gstudy" "list"
```

5.2.5 Decision (D) Study

The decision (D) study and generalizability (G) study, from which the generalizability and dependability coefficients can be estimated, were analyzed using the `gtheory` package (Moore, 2016).

```
decisionStudy <- dstudy(
  PxIxO,
  colname.objects = "Person",
  data = pio_cross_dat,
  colname.scores = "Score")
```

5.2.6 Generalizability Coefficient

The generalizability coefficient is analogous to the reliability coefficient in CTT. It divides the estimated person variance component (the universe score variance) by the estimated observed-score variance (with some adjustment for the number of observations). In other words, variance in a reliable measure should mostly be due to person variance rather than variance as a function of items, occasion, raters, methods, or other factors. The generalizability coefficient uses relative error variance, so it characterizes the similarity in the relative standing of individuals, similar to CTT-based estimates of relative reliability, such as Cronbach's alpha. Thus, the generalizability coefficient is an index of relative reliability. The generalizability coefficient ranges from 0–1, and higher scores reflect better reliability.

```
decisionStudy$generalizability
```

```
[1] 0.8731
```

5.2.7 Dependability Coefficient

The dependability coefficient is similar to the generalizability coefficient; however, it uses absolute error variance rather than relative error variance in the estimation. The dependability coefficient characterizes the absolute magnitude of differences across scores, not (just) the relative standing of individuals. Thus, the dependability coefficient is an index of absolute reliability. The dependability coefficient ranges from 0–1, and higher scores reflect better reliability.

```
decisionStudy$dependability
```

```
[1] 0.6374
```

5.3 Conclusion

G-theory provides an important reminder that reliability is not one thing. You cannot just say that a test "is reliable"; it is important to specify the facets across which the reliability and validity of a measure have been established (e.g., times, raters, items, groups, instruments). Generalizability theory can be a useful way of estimating multiple aspects of reliability and validity of measures in the same model.

5.4 Suggested Readings

Brennan (2001)