Generalizability Theory

I. Review of Basic Concepts of G Theory

II. Definitions of G Theory-Related Terms

III. CTT vs. GT

IV. Relation to ANOVA

V. Relative Coefficient of Generalizability

VI. Intraclass Correlation Coefficient

VII. Advantages of Generalizability Theory

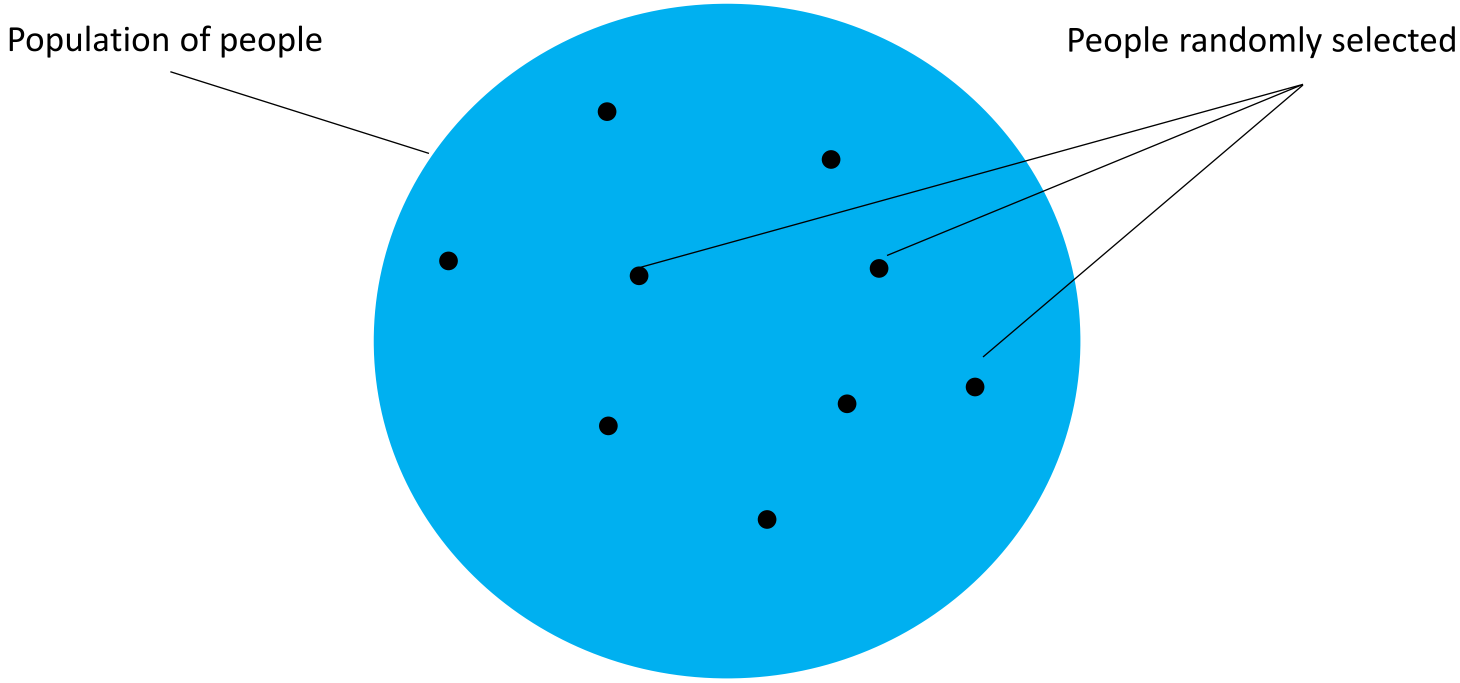
I. Review of Basic Concepts of G Theory

Recall that the *domain sampling model* (Cronbach, 1951; Tryon, 1957) states that if we can view each item as good representations of the true score and each as a random selected item from a domain or population of possible items, then we can relax the assumption that each test is strictly parallel in estimating reliability

Instead we only need to think of them as on average equally representing the domain

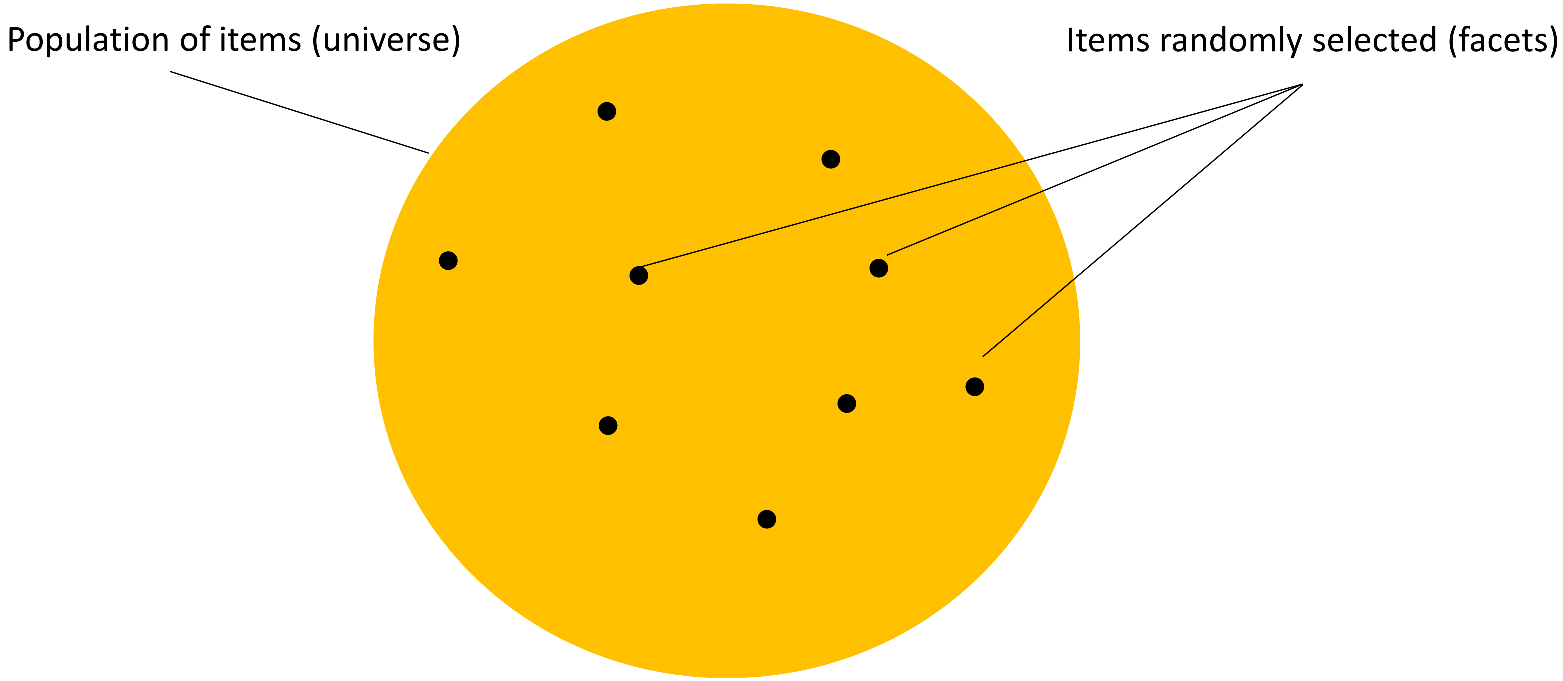
I. Review of Basic Concepts of G Theory

Sample of people from a population



I. Review of Basic Concepts of G Theory

Sample of items from a population



II. Definitions of G Theory-Related Terms

Generalizability theory (Cronbach, Gleser, Rajaratnam, 1963; Cronbach, Nageswari, & Gleser, 1963*)* is built on this domain sampling notion but goes beyond just items on a scale

Any type of observation can be used, each type is called a *facet*

- Ratings of behaviors (e.g., student problems) by different raters
- Observations of multiple events (e.g., vocalizations) related to a domain
- Different written profiles representing a domain (e.g., attractiveness profiles)
- Multiple time points (e.g., test, re-test), with each time point of measurement representing a different facet

There can be more than one facet (method/type of observation of the same construct) in a study

II. Definitions of G Theory-Related Terms

The *G study* is the generalizability phase of the research focusing on reliability or the extent to which the items generalizes the population of items

- Usually uses multiple facets to provide the best estimate of universe of scores

The *D study* is the decision phase of the research focusing on how the measure can be optimized for use in comparing groups or prediction

- Usually selects a facet (or a subset of facets) for a particular research question while minimizing error as much as possible

II. Definitions of G Theory-Related Terms

Items come from population of possible items, the *universe*

The full universe represents the true scores, so average of the universe of scores is the true score

The deviations from the universe score for a person are random instances

As might be expected, more items should more closely approximate the universe of items

III. CTT vs. GT

Recall classical test theory notion that each observed score is comprised of true score plus measurement error

$$\text{Reliability} = \frac{\text{True}}{\text{True} + \text{Error}}$$

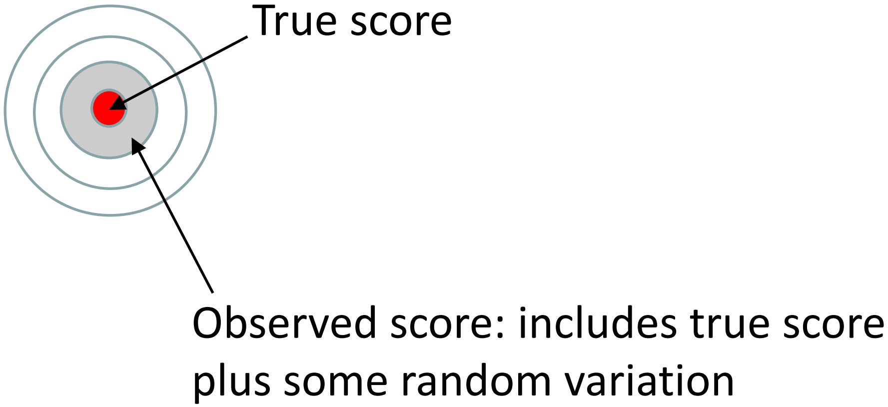
$$R_{xx} = \frac{s_t^2}{s_t^2 + s_e^2}$$

$$= \frac{s_t^2}{s_o^2}$$

Note: your text uses R_{xx} as the symbol for reliability but most texts use ρ_{xx} (rho) or r_{xx}

III. CTT vs. GT

Recall classical test theory notion that each observed score is comprised of true score plus measurement error



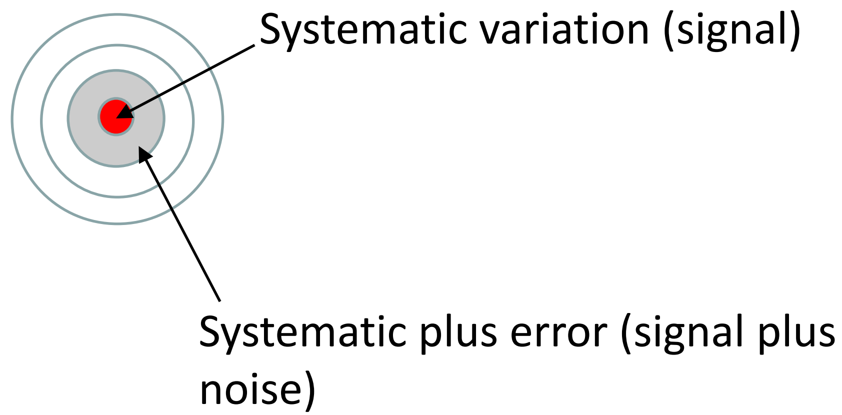
III. CTT vs. GT

The general notion is of systematic variation relative to random variation

$$\text{Reliability} = \frac{\text{True}}{\text{True} + \text{Error}} = \frac{\text{Signal}}{\text{Signal} + \text{Noise}} = \frac{\text{Systematic}}{\text{Systematic} + \text{Error}}$$

III. CTT vs. GT

For GT, variability of item values around the universe score is added noise but due to random variation of sampling of items



III. CTT vs. GT

For one facet, two types of systematic variation can be distinguished

Variation across individuals, averaging items, because some respondents rater higher or lower overall

Variation across items, averaging respondents, because some items are rated higher than other on average

Remaining “residual” variance represents error, capturing the random variation from sampling the domain

III. CTT vs. GT

In the end, generalizability theory is identical to classical test theory for a single facet

Just a different, more general way of conceptualizing reliability

IV. Relation to ANOVA

Generalizability and the notions of systematic and error variation also appear in analysis of variance (ANOVA)

ANOVA is a statistical test of two or more group means

Ronald Fisher proposed a method of testing differences among experimental groups by separating out between group variance from within-group variation

The differences between groups are due to the experimental conditions and therefore are due to systematic variation

IV. Relation to ANOVA

(Between-subjects design)

Group 1	Group 2	Group 3
5	1	7
4	3	6
2	4	4
5	2	6
7	2	6

$\bar{Y}_1 = 4.6$ $\bar{Y}_1 = 2.4$ $\bar{Y}_1 = 5.8$

Variation Within Group

Variation Between Groups

IV. Relation to ANOVA

When treatment is repeated measures, individual variation can be distinguished from error variation

Condition 1	Condition 2	Condition 3	
5	1	7	$\bar{Y}_1 = 4.3$
4	3	6	$\bar{Y}_2 = 4.3$
2	4	4	$\bar{Y}_3 = 3.0$
5	2	6	$\bar{Y}_4 = 4.3$
7	2	6	$\bar{Y}_5 = 5.0$
$\bar{Y}_1 = 4.6$	$\bar{Y}_1 = 2.4$	$\bar{Y}_1 = 5.8$	

Between-Individual Variation

Variation Between Time Points

IV. Relation to ANOVA

Each analysis, G theory, CTT, or ANOVA, uses the same general notion of partition or decomposition of the total observed variance into systematic and error variance (*variance components*)

V. Relative Coefficient of Generalizability

The variance components idea is used to compute an estimate of the generalizability of a measure

$$\text{Generalizability} = \frac{\text{Signal}}{\text{Signal} + \text{Noise}} = \frac{\text{Systematic}}{\text{Systematic} + \text{Error}}$$

To the extent that there is more systematic variation, there is greater generalizability, and the sample of items better represent the universe of items

V. Relative Coefficient of Generalizability

In the context of measurement and generalizability theory, we can consider variation across the respondents or *targets* of measurement and variation across items

Target variance, σ_t^2, represents the systematic variance and

Item variance, σ_i^2, represents how the items systematically vary across items (i.e., some items have higher values than others)

Residual variance, σ_{Res}^2, is the remaining variance not due to systematic variation across individuals or items

V. Relative Coefficient of Generalizability

Table 13.3 Equations for Estimating Variance Components in the Target Item Model

Effect	***Equation***
Target	$\sigma_t^2 = \dfrac{MS_t - MS_{Res}}{n_i}$
Item	$\sigma_i^2 = \dfrac{MS_i - MS_{Res}}{n_t}$
Residual	$\sigma_{Res}^2 = MS_{Res}$

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: an introduction*. Sage.

V. Relative Coefficient of Generalizability

$$\rho_t^2 = \frac{\sigma_t^2}{\sigma_t^2 + \dfrac{\sigma_{Res}^2}{n_i'}}$$

n_i' is the number of item (used or planned)

Notice that the larger between-target variance, the higher the generalizability coefficient

And that more items lead the term on the bottom to be smaller, so also increases the generalizability coefficient's value

V. Relative Coefficient of Generalizability

The relative coefficient of generalizability is the same as Cronbach's alpha (if only one facet is involved)

The *absolute generalizability* is a variant, used less often, that is used in criterion reference applications (e.g., cutoffs for hiring)

Item variation is added to the denominator, making absolute generalizability coefficients smaller than relative generalizability coefficients

VI. Intraclass Correlation Coefficient

Reliability can be evaluated when multiple raters observe and assess the same behavior—inter-rater reliability

e.g., three raters rate the aggressiveness of a child's behavior in the classroom

The Pearson correlation is one method of assessing inter-rater reliability

The average interrater correlation among a set of raters is the same as the relative coefficient of generalizability

VI. Intraclass Correlation Coefficient

The *intraclass correlation coefficient* (ICC) is often used to evaluate the correspondence between raters, particularly when the same raters do not all rate all the same participants

From a generalizability theory perspective, ratings from each rater are like multiple items and the participant is the target

VI. Intraclass Correlation Coefficient

There are several forms of the ICC but they all conceptually represent a ratio of between- and within-person variance

$$ICC = \frac{\text{between}}{\text{between} + \text{within}}$$

Some ICC coefficients for inter-rater reliability, ICC(2,1) and ICC(2,J) are equivalent to the corresponding absolute generalizability coefficients (Fan & Sun, 2014)

VI. Intraclass Correlation Coefficient

Imagine three raters of aggressive behavior of 8 pre-school children

Child	Rater 1	Rater 2	Rater 3	
1	**1**	**2**	**0**	
2	**1**	**3**	**3**	
3	**3**	**8**	**1**	
4	**6**	**4**	**3**	
5	**6**	**5**	**5**	
6	**7**	**5**	**6**	
7	**8**	**7**	**7**	
8	**9**	**9**	**9**	

VI. Intraclass Correlation Coefficient

Child	Rater 1	Rater 2	Rater 3	Mean
1	**1**	**2**	**0**	1.00
2	**1**	**3**	**3**	2.33
3	**3**	**8**	**1**	4.00
4	**6**	**4**	**3**	4.33
5	**6**	**5**	**5**	5.33
6	**7**	**5**	**6**	6.00
7	**8**	**7**	**7**	7.33
8	**9**	**9**	**9**	9.00
Mean	5.13	5.38	4.25	

Variation among children

VI. Intraclass Correlation Coefficient

Child	Rater 1	Rater 2	Rater 3	Mean
1	**1**	**2**	**0**	1.00
2	**1**	**3**	**3**	2.33
3	**3**	**8**	**1**	4.00
4	**6**	**4**	**3**	4.33
5	**6**	**5**	**5**	5.33
6	**7**	**5**	**6**	6.00
7	**8**	**7**	**7**	7.33
8	**9**	**9**	**9**	9.00
Mean	5.13	5.38	4.25	

Variation among raters

VI. Intraclass Correlation Coefficient

Child	Rater 1	Rater 2	Rater 3	Mean
1	**1**	**2**	**0**	1.00
2	**1**	**3**	**3**	2.33
3	**3**	**8**	**1**	4.00
4	**6**	**4**	**3**	4.33
5	**6**	**5**	**5**	5.33
6	**7**	**5**	**6**	6.00
7	**8**	**7**	**7**	7.33
8	**9**	**9**	**9**	9.00
Mean	5.13	5.38	4.25	

Variation among raters within children

VI. Intraclass Correlation Coefficient

$$ICC = \frac{\text{Variation among children}}{\text{Variation among children} + \text{Variation among raters} + \text{Variation among raters within children}}$$

VII. Advantages of Generalizability Theory

Unified, broadened framework that extends the classical test theory notion to observations other than items (e.g., ratings)

Extends reliability estimation to multiple facets (e.g., items, raters, occasions) which can also be investigated in combination

Allows for estimation of generalizability for rating designs that are not fully crossed (i.e., all raters to not need to rate all targets)