Promoting Practical Understanding of Generalizability Theory within School-Based Research

Northeastern University
Bouvé College *of* Health Sciences

Amy Briesch, Ph.D.[1] *Sandra M. Chafouleas, Ph.D.*[2]
Northeastern University[1] *University of Connecticut*[2]

CBER
Promoting Academic & Behavioral Supports
Center for Behavioral Education & Research
UCONN NEAG SCHOOL OF EDUCATION

Introduction

Generalizability Theory (GT; Cronbach et al.,1972) offers increased utility for school-based research given the ability to (1) inform both relative and absolute decision making, (2) concurrently examine multiple sources of variance, and (3) determine the generalizability of results (rather than simply consistency). Despite these strengths, exploration into its use and application has been conducted by relatively few researchers within the field of school psychology. In order to keep pace with the evolving assessment needs within service delivery models in education that emphasize accountability and prevention (e.g. RtI), there is a need to expand methodological skill and understanding specific to psychometric requirements. Although GT is not the panacea to answer all needs in educational assessment, its use can provide psychometric information that is likely to be more in line with the current questions facing the field of school psychology. This analytic approach has been underutilized among researchers engaged in assessment development and evaluation, however, and lack of guidance regarding appropriate use has resulted in potentially inappropriate or incomplete application.

Key methodological considerations

Before arriving at the stage of data analysis, several methodological and analytic considerations must be made. Considerations related to use of GT have been discussed (Smith, 1981) and desiderata for evaluating studies submitted for publication outlined (Hendrickson & Yin, 2010); however, integrated guidelines are provided below in order to facilitate understanding of the range of factors that need to be considered in the design and implementation of G and D studies.

Inclusion of relevant facets
•Define universe of admissible observations (acceptable measurement conditions)
Specification of facet classification
•Define the universe of generalization (i.e. conditions of a facet to which a decision maker wishes to generalize)
•Facets may be treated as either random or fixed, a decision that determines the extent to which results can be generalized to other measurement situations
Use of crossed versus nested designs
•Use of a fully-crossed design is desirable, in that it permits the researcher to interpret all facets and interactions. Unfortunately, however, use of a fully-crossed design is not always logistically feasible.
Balancing sample size with model complexity
•As the complexity of the model increases, so, too, does the degree of sampling error inherent in variance component estimates.
Selection of an appropriate computer program
•Variance components may be derived using a variety of estimation methods including ANOVA-like procedures, maximum likelihood (ML) procedures, or minimum norm quadratic unbiased estimation (MINQUE). Although several computer programs may be used to estimate these variance components (e.g., GENOVA, Crick & Brennan, 1993; SPSS, SPSS, 1997; SAS, SAS Institute, 1996), it is important to note that these programs utilize different estimation procedures, and therefore require different assumptions about the underlying data

For additional information, please direct all correspondence to Amy Briesch at a.briesch@neu.edu

Briesch, A. M., & Chafouleas, S. M. (2011, August). *Promoting practical understanding of Generalizability Theory within school-based research*. Poster presentation at the Annual Convention of the American Psychological Association, Washington DC.

Literature Review

In order to illustrate the range of applications of GT within the fields of education and psychology, a literature search was conducted using the PsycINFO and ERIC databases. Keywords used to conduct the search were: (a) generalizability, (b) reliability, (c) measurement, (d) psychometrics, and (e) variance. All peer-reviewed studies that utilized GT to examine school-based outcome data were included, regardless of the assessment domain (e.g. academic, behavioral). A total of 34 studies were identified between 1976 and 2010, with over two-thirds of these papers (*n* = 23) published within the past 10 years. Of these studies, the majority (*n* = 21) investigated the dependability of academic measures, such as Curriculum-Based Measurement probes (e.g., Hintze, Owen, Shapiro, & Daly, 2000) and rubric-based evaluations of student writing performance (e.g., Jiang & Smith, 2000). Behavioral assessments were also commonly examined (*n* = 11), including both observational (e.g., Coates & Thoresen, 1978) and rating-based data (e.g., Bergeron, Floyd, McCormack, & Farmer, 2008). Models were found to range in complexity from one-facet designs (e.g., Christ & Ardoin, 2009; Poncy et al., 2005) to designs involving nesting and multiple facets (e.g., Macready, 1983; Tindal et al., 2008). Details can be found in the table below.

Citation	Domain	Outcome Measure	Design
Bergeron, Floyd, McCormack, & Farmer (2008)	Behavior	Externalizing composite and subscales scores on: • Behavior Assessment System for Children-Second Edition (Teacher Rating Form) • ASEBA-Teacher Rating Form	((Students x Raters): Classrooms) x Occasions x Instruments
Briesch, Chafouleas, & Riley-Tillman (2010)	Behavior	Teacher ratings of Academic Engagement using: • Systematic direct observation • Direct Behavior Rating	Persons x Raters x (Occasions: Day)
Brown-Chidsey, Davis, & Maya (2003)	Academic	CBM-Silent Reading passage scores	Persons x Grade x Special Education Status
Chafouleas, Briesch, Riley-Tillman, Christ, Black, & Kilgus (2010)	Behavior	Teacher-completed DBR for middle school students: Academic Engagement, Disruptive Behavior	Persons x Raters x (Occasions: Day)
Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese (2007)	Behavior	Teacher-completed DBR for preschool students: Works to Resolve Conflicts, Interacts Cooperatively with Peers	Persons x Raters x Day x Setting
Christ & Ardoin (2009)	Academic	Correct words per minute on Curriculum-Based Measurement-Reading probes	Persons x Passages
Christ, Johnson-Gros, & Hintze (2005)	Academic	Scores on multiple-skill CBM-Math computation probes	Persons x Assessment Duration
Christ & Vining (2006)	Academic	Scores on multiple-skill CBM-Math probes with either random or stratified stimulus sets	Persons x Probes
Coates & Thoresen (1978)	Behavior	Observational data (Eating Analysis and Treatment Schedule)	Observers x (Times: Persons)
Fawson, Reutzel, Smith, Ludlow, & Sudweeks (2006)	Academic	Scores on running record assessments (Reading Recovery program)	(Students x Passages): Raters
Fitzpatrick, Lee, & Gao (2001)	Academic	School-level scores on short test forms containing open-ended, constructed-response items measuring mathematical skills	Persons: (Schools x Test Forms)
Gierl (1998)	Academic	Scores on written-response tasks from an English diploma examination	Persons x Rater x Scale (e.g. Organization, Writing Skills)
Hintze, Christ, & Keller (2002)	Academic	Scores on single- and multiple-skill CBM-Math probes	(Persons: Grade) x (Probe: Probe Type)
Hintze & Matthews (2004)	Behavior	Systematic direct observation (momentary-time sampling) of on- and off-task behavior	Persons x Time x Setting
Hintze, Owen, Shapiro, & Daly (2000)	Academic	Words correct per minute on (1) literature- and skills-based CBM-Reading passages, (2) instructional- and challenging-level CBM-Reading passages	(Persons: Grade) x Method x Occasion
Hintze, & Pelle Pettite (2001)	Academic	CWPM on grade-level ORF passages	(Persons: Special Education Status) x Occasion
Johnson & Bell (1985)	Academic	Responses to questions reflecting science knowledge	(Persons: (Schools x Gender): Forms)

Citation	Domain	Outcome Measure	Design
Kan (2007)	Academic	Scores on a Reading Comprehension task	Persons x Scoring Type (i.e. Guide/No Guide) x Rater
Lane, Liu, Ankenmann, & Stone (1996)	Academic	Scores on a measure of outcomes and growth in mathematics (QUASAR Cognitive Assessment Instrument)	Person x Task Persons x Task x Rater (Persons: School) x Task
Lee (2002)	Academic	Scores on complex reading comprehension tests	Multiple models involving Items, Passages, Content, Themes, Types of Passages
Lee & Fitzpatrick (2003)	Academic	Scores on a statewide mathematics assessment measuring computation and application skills	Persons: (School x Test Form)
Lomax (1982)	Behavior	Observational data (Student-Level Observation of Beginning Reading)	Persons x Rater x Observational Measure (i.e. category)
Macready (1983)	Academic	Scores on multiplication test consisting of three- and four-digit multiplicands	(Persons: Classrooms) x (Items: (Domains x Number of Digits))
Marcus (1980)	Behavior	Observational data (cooperative behavior) during preschool free play periods	Persons x Occasions
Martinez, Goldschmidt, Niemi, Baker, & Sylvester (2007)	Academic	Rubric-based scores on a prompted essay for the English-Language Arts Performance Assignment	Persons x (Raters: Districts)
McWilliam & Ware (1994)	Behavior	Observational data of student engagement	Persons x Occasions x Raters
Newton (2010)	Behavior	Observational data of teacher practice (qualitative data transformed to rubric scores)	Persons x Raters x Occasions
Poncy, Skinner, & Axtell (2005)	Academic	Words correct per minute on CBM-R ORF passages	Persons x Items
Smith & Kulikowich (2004)	Other	Scores on assessment designed to measure complex problem-solving skills (Kickball Assessment)	Persons x Items x Raters x Occasions
Suen, Lu, Neisworth, & Bagnato (1993)	Other	Scores from the System to Plan Early Childhood Services (SPECS) assessment procedure	(Raters: Persons) x Items
Swartz, Hooper, Montgomery, Wakely, de Kruif, et al. (1999)	Academic	Subtest scores on Test of Written Language-2 (TOWL-2)	Persons x Raters
Swartz et al. (1999).	Academic	Holistic scoring rubric developed by NAEP; Analytic scoring rubric developed by researchers	Persons x Raters
Tindal, Yovanoff, & Geller (2008)	Academic	Scores on alternate reading assessment	Persons x Items: (Administrative Format x Task) x Raters
Volpe, McConaughy, & Hintze (2009)	Behavior	Scale scores from the ASEBA Direct Observation Form	Persons x (Occasions: Time of Day)
Zhang, Johnston, & Kilic (2008)	Behavior	Self and peer rubric-based ratings from group work	(Persons x Raters):Group

Glossary

Generalizability study = estimation of variance components

Dependability study = variance components used to derive reliability-like coefficients
•Generalizability (G) coefficient = used to inform relative (i.e. inter-individual) decision making
•Dependability (D) coefficient = used to inform absolute (i.e. intra-individual) decision making

Facets ≈ factors; any set of conditions under which measurements can be carried out; a possible source of measurement error

Universe of generalization = conditions of a facet to which a decision maker wishes to generalize

Universe of admissible observations = all possible observations (e.g., raters, days) deemed acceptable to the decision maker

Fixed versus Random = decision made by investigator for each facet regarding whether conditions in the D study are a sample from those in the universe of generalization (random) or are exhausted (fixed). Note that each decision has implications – for example, "fixing" tends to lower error variance and increase coefficients.