Journal of School Psychology 52 (2014) 13–35

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Commissioned Article

Generalizability theory: A practical guide to study design, implementation, and interpretation☆

Amy M. Briesch [a,*], Hariharan Swaminathan [b], Megan Welsh [b], Sandra M. Chafouleas [b]

[a] *Northeastern University, USA*
[b] *University of Connecticut, USA*

ARTICLE INFO

Article history:
Received 24 June 2013
Received in revised form 29 November 2013
Accepted 30 November 2013

Keywords:
Generalizability theory
Dependability
Measurement
School-based assessment

ABSTRACT

Generalizability Theory (GT) offers increased utility for assessment research given the ability to concurrently examine multiple sources of variance, inform both relative and absolute decision making, and determine both the consistency and generalizability of results. Despite these strengths, assessment researchers within the fields of education and psychology have been slow to adopt and utilize a GT approach. This underutilization may be due to an incomplete understanding of the conceptual underpinnings of GT, the actual steps involved in designing and implementing generalizability studies, or some combination of both issues. The goal of the current article is therefore two-fold: (a) to provide readers with the conceptual background and terminology related to the use of GT and (b) to facilitate understanding of the range of issues that need to be considered in the design, implementation, and interpretation of generalizability and dependability studies. Given the relevance of this analytic approach to applied assessment contexts, there exists a need to ensure that GT is both accessible to, and understood by, researchers in education and psychology. Important methodological and analytical considerations are presented and implications for applied use are described.

> *There is no single, universal, and absolute reliability coefficient for a test. Determination of reliability is as much a logical as a statistical problem.*
>
> [Thorndike, 1951, pp. 570–571]

1. Introduction

Humans have been interested in the measurement of individual differences ever since the days in which Sir Francis Galton conducted measurements of sensory (e.g., auditory and tactile) discrimination abilities in his Anthropometric Laboratory (Jones & Thissen, 2007). Over the past century, however, it is clear that the nature of psychometric assessment has changed substantially. First, few of the constructs (i.e., traits or concepts) that are of most interest to educators and psychologists can be measured using precise instruments, such as a heart rate monitor; rather, tests and procedures must be developed in order to measure unobservable phenomena such as abilities, attitudes, and personality traits. Because it is not possible to directly measure constructs as is true in

☆ The authors would like to thank editors Ed Daly and Randy Floyd for their enthusiastic support throughout preparation of the manuscript as well as the many colleagues who have facilitated extension of G theory into educational and psychological assessment.

* Corresponding author at: Northeastern University, Department of Counseling and Applied Educational Psychology, 360 Huntington Avenue, Boston, MA 02115, USA.

E-mail address: a.briesch@neu.edu (A.M. Briesch).

ACTION EDITOR: Randy Floyd.

0022-4405/$ – see front matter © 2013 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jsp.2013.11.008

biological or physical measurement, greater attention must be paid to the identification and regulation of measurement variance. Second, assessment contexts have changed over time from carefully-controlled laboratories to applied settings such as schools and communities. Given that it is often necessary in applied assessment contexts to measure behavior using different raters or different forms or at different times, multiple potential sources of measurement error must be both identified and examined. Third, summative, criterion-based assessments are no longer considered to be sufficient to answer the questions posed by both researchers and practitioners. As one example, models of educational service delivery that emphasize early identification and intervention have necessitated a shift in assessment orientation toward formative purposes (i.e., initial screening and then progress monitoring). As new measures emerge to meet changing assessment purposes, new challenges are posed around the evaluation of psychometric evidence. Taken together, there is a need to expand methodological skill and understanding specific to psychometric requirements.

Shifts in the assessment landscape necessitate new ways of thinking about psychometric evidence that extend beyond options available in classical test theory (CTT). Generalizability Theory (GT; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) offers distinct advantages when applied in contemporary assessment contexts including the fact that it is possible to examine multiple sources of variance simultaneously and examine both the consistency and generalizability (i.e., representativeness of a specific sample of behavior) of measurement. However, with the exception of those researchers and test developers working within the context of large-scale performance assessments (e.g., Gao, Shavelson, & Baxter, 1994), full exploration into the use and application of GT has only recently emerged. GT will only help to fill a gap if clear and appropriate guidance regarding its use is available. Therefore, the purpose of this article is to provide an overview of GT, focusing on methodological and analytic concerns, in order to facilitate understanding and use of GT by researchers in education and psychology.

2. Theoretical background and fundamental concepts

Although individuals may have different views in terms of what constitutes psychometric adequacy, most people can agree that a measurement is only useful to the extent that it provides meaningful information about individuals. Regardless of whether we are interested in assessing a child's ability to read or an individual's self-concept, we are not interested in what a specific sample of behavior looks like at a specific point in time but rather in assessing an underlying trait or characteristic. Plato suggested that each object or quality that humans come across in their day-to-day lives is merely a shadow or representation of a true Form (Ley, 1972). Similar to the idea of a Platonic Form, we can never know the exact value of a trait or characteristic. Estimation of this true value is, however, the focus of CTT approaches.

One of the central assumptions of a CTT approach is that an observed score (X) is composed of two components: a *true score* (T) and some degree of *measurement error* (e). Although it is not possible to directly measure the true score, T can be estimated by determining the average score obtained across the administration of a hypothetically infinite number of parallel measurements (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Even when using parallel measurements, however, it is accepted that an individual's behavior can never be perfectly consistent from one sample to another due to both internal (e.g., changes in motivation or health) and external factors (e.g., changes in testing conditions or rating severity). Although each observed score represents an attempt to estimate the true score, each observed score is also accompanied by some degree of random and unpredictable error. The difference between one's hypothetical true score and the observed score is therefore what is termed measurement error. Although e could represent many different types of error (i.e., rater error or instrument error), CTT only permits the estimation of one general error term, which incorporates all sources of error but emphasizes certain types of error over others depending on the reliability study conducted (e.g., interrater reliability focusing on rater error and test–retest reliability focusing on temporal error). Several different types of reliability coefficients can be estimated; however, each is defined as the ratio of true score variance to observed (i.e., true plus error) score variance (Crocker & Algina, 2006).

When calculating traditional reliability coefficients within a CTT framework, the goal is to determine the degree to which variations in the conditions of measurement (e.g., different raters and different measurement occasions) affect the consistency with which a construct is measured. In the early years of psychological measurement, when the target constructs consisted of precise, tightly-controlled measurements (e.g., reaction time and eye movements), and measurements were made in controlled laboratory environments, CTT metrics were capable of providing psychometric data that were easily interpretable and had direct relevance to the questions posed because the potential sources of error were limited. As the scope of psychological measurement expanded, however, to include a wide range of human behaviors that are inarguably influenced by a host of different factors such as time and setting, it has been suggested by some that the use of classic psychometric concepts may not be most appropriate (e.g., Nelson, Hay, & Hay, 1977). This criticism has particular relevance to educational and psychological assessment, given the degree of variability typically encountered within applied settings. In real-world assessment, where the players, settings, and even foci of assessment are ever-changing, a traditional assessment approach (i.e., one that estimates one global error term) has limited utility. Taken together, the expanding purposes for educational and psychological assessments and guidelines for evaluation of tools to meet the changing needs suggest that complements to CTT are needed.

3. Key concepts underlying the use of generalizability theory

Within a CTT framework, reliability coefficients are commonly computed using the Pearson product–moment correlation; however, a Pearson correlation coefficient is only useful where we can identify the X and Y scores in the pair (X and Y). Karl Pearson therefore introduced the intraclass correlation (ICC) around the turn of the 20th century to assess relations among