Generalizing from Educational Research

Beyond Qualitative and Quantitative Polarization

EDITED BY
Kadriye Ercikan Wolff-Michael Roth

Generalizing from Educational Research

> "… a potent force for change in the field. This volume is ground-breaking, with the potential to make a major paradigm shift in thinking about educational research. It is the kind of volume that should be on the shelves of virtually everyone doing research in education."
>
> *Peter Seixas, University of British Columbia*

> "This book frames the major challenge facing educational researchers as one of going beyond the mindless qualitative-quantitative divide and addressing the overarching/fundamental challenge of enriching and enlarging educational inquiry. It is a signature contribution to the field."
>
> *Clifton F. Conrad, University of Wisconsin-Madison*

Tackling one of the most critical issues in education research today—how research methods are related to value and meaningfulness—this frontline volume achieves two purposes. First, it presents an integrated approach to educational inquiry that works toward a continuum instead of a dichotomy of generalizability, and looks at how this continuum might be related to types of research questions asked and how these questions should determine modes of inquiry. Second, it discusses and demonstrates the contributions of different data types and modes of research to generalizability of research findings, and to limitations of research findings that utilize a single approach.

International leaders in the field take the discussion of generalizing in education research to a level where claims are supported using multiple types of evidence. This volume pushes the field in a different direction, where the focus is on creating meaningful research findings that are not polarized by qualitative versus quantitative methodologies. The integrative approach allows readers to better understand possibilities and shortcomings of different types of research.

Kadriye Ercikan is Associate Professor of Measurement and Research Methods in the Department of Educational and Counseling Psychology and Special Education, at the University of British Columbia, Canada.

Wolff-Michael Roth is Lansdowne Professor of Applied Cognitive Science at the University of Victoria, Canada.

Generalizing from Educational Research

Beyond Qualitative and Quantitative Polarization

Edited by

Kadriye Ercikan
University of British Columbia, Canada

and

Wolff-Michael Roth
University of Victoria, Canada

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2009
by Routledge
270 Madison Ave, New York, NY 10016

Simultaneously published in the UK
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

This edition published in the Taylor & Francis e-Library, 2008.

"To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk."

Library of Congress Cataloging in Publication Data
Generalizing from educational research / edited by Kadriye Ercikan and Wolff-Michael Roth.
p. cm.
Includes bibliographical references and index.
1. Education—Research—Methodology. I. Ercikan, Kadriye. II. Roth, Wolff-Michael, 1953–
LB1028.G398 2008
370.72—dc22
2008027765

ISBN 0-203-88537-6 Master e-book ISBN

ISBN 10: 0-415-96381-8 (hbk)
ISBN 10: 0-415-96382-6 (pbk)
ISBN 10: 0-203-88537-6 (ebk)

ISBN 13: 978-0-415-96381-7 (hbk)
ISBN 13: 978-0-415-96382-4 (pbk)
ISBN 13: 978-0-203-88537-6 (ebk)

Contents

Preface ix

1 Introduction 1
WOLFF-MICHAEL ROTH AND KADRIYE ERCIKAN

SECTION I

Generalizing Within and Beyond Populations and Contexts **9**

Overview

2 Generalizability Theory and Its Contribution to the Discussion of the Generalizability of Research Findings 13
RICHARD J. SHAVELSON AND NOREEN M. WEBB

3 The Testing of English Language Learners as a Stochastic Process: Population Misspecification, Measurement Error, and Overgeneralization **33**
GUILLERMO SOLANO-FLORES

Section I Highlights 46

SECTION II

Combining and Contrasting Qualitative and Quantitative Evidence **49**

Overview

4 Generalization from Qualitative Inquiry **51**
MARGARET EISENHART

5 **On Qualitative and Quantitative Reasoning in Validity** 67
ROBERT J. MISLEVY, PAMELA A. MOSS, AND JAMES P. GEE

6 **Generalizability and Research Synthesis** 101
BETSY J. BECKER AND MENG-JIA WU

Section II Highlights 117

SECTION III

How Research Use Mediates Generalization 123

Overview

7 **Generalizability and Research Use Arguments** 127
LYLE F. BACHMAN

8 **Repetition, Difference, and Rising Up with Research in Education** 149
KENNETH TOBIN

9 **Critical Realism, Policy, and Educational Research** 173
ALLAN LUKE

Section III Highlights 201

SECTION IV

Rethinking the Relationship Between the General and the Particular 207

Overview

10 **Limitations in Sample-to-Population Generalizing** 211
KADRIYE ERCIKAN

11 **Phenomenological and Dialectical Perspectives on the Relation between the General and the Particular** 235
WOLFF-MICHAEL ROTH

Section IV Highlights 261

12 Discussion of Key Issues in Generalizing in Educational Research 265
EDITED BY KADRIYE ERCIKAN AND WOLFF-MICHAEL ROTH WITH CONTRIBUTIONS FROM LYLE F. BACHMAN, MARGARET EISENHART, ROBERT J. MISLEVY, PAMELA A. MOSS, GUILLERMO SOLANO-FLORES, AND KENNETH TOBIN

Contributors 295
Index 299

Preface

In 2007, we (the editors) invited some colleagues who had been working, explicitly or implicitly, on the problem of generalizing in educational research to contribute to a book on the topic. These are colleagues who look at assessments as evidentiary reasoning and at validity from different perspectives including situative and hermeneutic, who discuss utility of information from research as an integral part of validity of generalizing from small-scale studies, or who think about the validity of assessment interpretations for special populations such as English language learners. For these colleagues, generalizability theory as a measurement theory deals explicitly with such factors and the degree to which they affect generalizing from assessment results. Their contributions to this volume focus on different forms of generalizing, for example, across individuals versus across time or across several research studies, such as in the case of meta-analysis; or they focus on the relationship between the particular and the general.

In 2007, when we put together a proposal for a symposium for the American Educational Research Association Annual Meeting on the issues discussed in this book, even though we believed the topic would be of interest to all educational researchers, Division D (Measurement and Research Methodology) seemed to be the most appropriate to which to submit our proposal. The next step in the submission process was to decide to which section of Division D to submit it. The point here is that our proposal submission faced the polarization and boundaries that currently exist in conceptualizations of research methods and in our disciplines in general. Division D has three sections, one dedicated to measurement, another to quantitative research, and a third to qualitative research. Our proposal cut across interests of all three sections, whereas one of the primary goals of the symposium—and this book—was to break down boundaries such as those that currently exist in Division D. Similar boundaries exist in our journal foci and in specializations in the association of generalization (or lack thereof) to types of data and data

summarization techniques. These boundaries do not serve education or education research well. Our intention in this book is to move education research in the direction of prioritizing research questions and knowledge creation using many and multiple modes of research and go beyond simplistic divisions of research types.

Chapter 1

Introduction

Wolff-Michael Roth and Kadriye Ercikan

A fundamental presupposition in cultural-historical activity theory—an increasingly used framework for understanding complex practical activities, interventions, and technologies in schools and other workplaces (e.g., Roth & Lee, 2007)—is that one cannot understand some social situation or product of labor without taking into account the history of the culture that gave rise to the phenomenon. This book, too, can be understood only within the particulars of some fortunate events that brought us, the two editors, together in quite unpredictable ways. From this coming together emerged a series of collaborative efforts in which we, normally concerned with teaching quantitative and qualitative methods, respectively, began to think about educational research more broadly and transcending the traditional boundaries around doing inferential statistics and designing (quasi-) experiments and doing various forms of naturalistic inquiry.

Some time during the early part of 2004, King Beach and Betsy Becker approached the two of us independently inviting us to co-author a chapter for a section in a handbook of educational research that they edited. Kadriye works in the program area of measurement, evaluation, and research methodology and focuses on psychometric issues; Michael, though trained as a statistician, teaches courses in interpretive inquiry. Despite or perhaps because of the apparent differences, we both tentatively agreed and, soon thereafter, met when Michael participated at a conference in Vancouver, the city where Kadriye's university is located. During the subsequent months, we interacted both on the phone, via e-mail, and through our mutual engagement with each other's texts and contributions to the joint endeavor. The collaboration as process and our final product on "Constructing Data" (Ercikan & Roth, 2006a) both provided us with such a great satisfaction that we soon thereafter decided to work together on another project, this time dealing with one of the issues that had emerged from our collaboration: the apparent opposition of "quantitative" and "qualitative" approaches in educational research, which in the past has lead to insurmountable conflicts

and paradigm wars. We, however, had felt while writing the handbook chapter that there are possibilities for individuals such as ourselves with different research agendas and methods to interact collegially and productively. We decided to grabble with the question, "What good is polarizing research into qualitative and quantitative?" and to report the outcome of our collaborative investigation to a large audience of educators (Ercikan & Roth, 2006b).

In the process of working on the question, we came to argue against polarizing research into quantitative and qualitative categories or into the associated categories of subjective and objective forms of research. We demonstrated that this polarization is not meaningful or productive for our enterprise, educational research. We then proposed an integrated approach to education research inquiry that occurs along a continuum instead of a dichotomy of generalizability. We suggested that this continuum of generalizability may be a function of the types of research questions asked; and it is these questions that ought to determine the modes of inquiry rather than any a priori questions about the modes of inquiry—which drives the "monomaniacs" (Bourdieu, 1992) of method, which build entire schools and research traditions around one technique.

As during our first collaborative venture, we emerged from this experience both satisfied to have conducted a conversation across what often is perceived to be a grand divide and to have achieved a worthwhile result. Not soon after completion, we began talking about a special issue in a journal that would assemble leading scholars in the field discussing the issues surrounding generalization in and from educational research. But it became quite clear in our early discussions that the format of a journal issue would be limiting the number of people we could involve and the formats that the individual pieces could take. It also would limit us in producing the kind of coherent work that we present in this book, where chapters are bundled into sections, with an all encompassing narrative that provides linkages between chapters and starting points for further discussion. Our motive for this book, elaborated in the following section, was to have our contributors think about questions arising for them in the endeavor to generalize from educational research with the aim of going beyond the dichotomous opposition of quantitative and qualitative research method.

Beyond the Quantitative–Qualitative Oppositions

The discussion concerning the generalizability was sharpened with and in the debate between what came to be two camps, those doing statistics and denoting their work as "quantitative" and those doing other forms of inquiry denoted by the term "qualitative" or "naturalistic."

The discussion was polarized, among others, in Yvonna Lincoln and Egon Guba's (1985) *Naturalistic Inquiry*, where the "naturalist paradigm" was presented to be the polar opposite to "positivist paradigm." Accordingly, naturalists were said to have difficulties with the concept of external validity, that is, the generalizability of research-generated knowledge beyond the context of its application. The *transferability* of findings made in one context to some other context was taken to be an empirical matter rather than one that could be assumed based on statistical inference, even with its safeguards of estimating the probability of type I and type II errors. The classical position assumed that given high internal validity in some sample A and given the sample is representative of the population P, then findings made in the sample A could be generalized to the population P as a whole, and, therefore, to all other samples that might be taken from it.

The so-called naturalists rejected this form of generalization. One of the main points of the rejection is grounded in the very idea of a population. Guba and Lincoln remind their readers that inferences about populations can be improved with the specification of "homogeneous strata." But this in fact constitutes a specification of context and contextualization of knowledge. This therefore raises the issue about the extent to which something found in some inner-city school district in Miami can be used to inform teaching and learning in inner-city Philadelphia or New York, i.e., the three cities where one of our chapter authors, Kenneth Tobin, has conducted detailed studies of teaching and learning science. Concerning teaching, we know from detailed ethnographic work that a Cuban-African American science teacher highly successful in inner-city Miami was unsuccessful in his own account teaching science to "equivalent" students in inner-city Philadelphia. But the same teacher, much more quickly than other (novice) teachers, became a highly effective teacher in this for his new environment. Thus, his practical knowledge of teaching science to disadvantaged students turned out to be both transferable and non-transferable.

The discussion concerning the generalizability of educational research in the United States has heated up again during the George W. Bush era, when policy makers declared that experimental design constituted the "gold standard" of (educational) research. All other forms of research generally and "qualitative research" more specifically, were denigrated as inferior. In this book, we invited well-established and renowned researchers across the entire spectrum of educational research methods to weigh in on the question concerning the extent to which educational research can be generalized and transported (transferred) to other contexts.

Generalization and generalizability are gaining more importance with increased levels of scrutiny of value and utility of different types

of educational research by funding agencies, the public, educational community and researchers themselves. These aspects of educational research have come to define the utility and quality of research in education and have also come to further polarize conceptualizations of educational research methods (Shaffer & Serlin, 2004). In light of the present political debates about the usefulness of different kinds of research (e.g., the "gold standard"), the issue of generalizability is often entered into the discussion as a criterion to argue for one form of research as superior over another. Typically, the scholarly (and political) discussion of degrees of *generalizability* is inherently associated with statistical (i.e., "quantitative") approaches and commonly questions the generalizability of observational (i.e., "qualitative") approaches. Unlike often assumed, we argued in our *Educational Researcher* feature article that a quantitative–qualitative distinction does not correspond to a distinction of the presence and absence of generalizability (Ercikan & Roth, 2006b). Rather, there are "qualitative" forms of research with high levels of generalizability and there are "quantitative" forms of research with rather low levels of generalizability. In addition, we argued and demonstrated that research limited to polar ends of a continuum of a variety of research methods, such as experimental design in evaluating effectiveness of intervention programs, in fact can have critically limited generalizability to decision making about sub-groups or individuals in intervention contexts.

One of the central issues may be the usefulness of different types of data and descriptions useful to different stakeholders in the educational enterprise. Thus, the following graphical representation that a researcher may have constructed to correlate the performance on a pre-test with scores indicating a particular learning outcome. Whereas the pretest might be consistent with published research and therefore reproduce

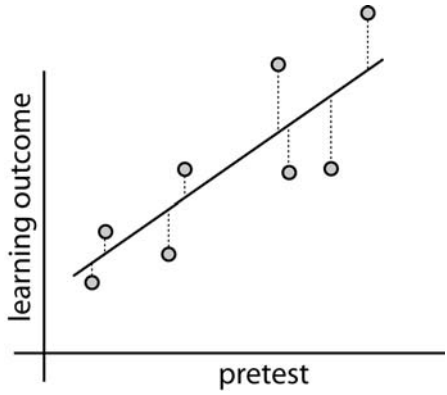

Figure 1.1 Correlation between pretest and learning outcomes.

existing (statistically reliable) relationships with the learning outcome variable, knowing the correlation actually helps a classroom teacher very little. The teacher, to design appropriate instruction for individual students, is interested precisely in the variation from the trend, that is, she is interested in the variation that in statistical approaches constitutes error variance. That is, to properly inform this teacher on what to do in her classroom, we need to provide her with forms of knowledge that are simultaneously sufficiently general to provide her with trends and with forms of knowledge that are sufficiently specific to allow her to design instructions to the specific needs expressed in the variation from the trend.

This book is designed to address these issues in a comprehensive way, drawing on the expertise of leading, well-known researchers in the fields traditionally characterized by the adjectives qualitative and quantitative research. The purpose of this book is twofold: (a) to work out and present an integrated approach to educational research inquiry by aiming at a continuum instead of a dichotomy of generalizability, how this continuum might be related to types of research questions asked and how these questions should determine modes of inquiry; (b) to discuss and demonstrate contributions of different data types, and modes of research to generalizability of research findings and limitations of research findings in research that utilizes a single research approach.

Arguing against single-method research but for generalization, Pierre Bourdieu (1992) portrays analogical reasoning to be one of the powerful instruments of research. Analogical reasoning allows researchers to immerse themselves in the particularities of their cases without drowning in them—a familiar experience to many novice researchers. As Bourdieu elaborates, analogical reasoning realizes generalization

> not through the extraneous and artificial application of formal and empty conceptual constructions, but through this particular manner of thinking the particular case which consists of actually thinking it as such. This mode of thinking fully accomplishes itself logically in and through the comparative method that allows you to think relationally a particular case constitutes as a "particular instance of the possible" by resting on the structural homologies that exist between different fields … or between different states of the same field. (p. 234)

In the course of this book, we work toward such a conception of generalization in educational research, as outlined in more or less the same form in the chapter by Wolff-Michael Roth, who takes a similar dialectical perspective as Bourdieu though grounded in and arising from a different scholarly context.

Most importantly, this book not only is about generalizing from educational research but also is and arose from the self-questioning accomplished researchers engaged in when we asked them to address the question at the heart of this book. We emerge from this work with a sense that there is a lot of recognition for the different problems arising from different forms of inquiry, a mutual respect, and a desire to continue to contribute to resolving the hard question: how to make research relevant to all stakeholders in the educational enterprise.

Structure and Content

This book consists of 11 chapters clustered into four sections: "Generalizing Within and Beyond Populations and Contexts," "Combining and Contrasting Qualitative and Quantitative Evidence," "How Research Use Mediates Generalization," and "Rethinking the Relationship Between the General and the Particular." Each section begins with an overview text presenting and contextualizing the main ideas that gather the chapters in the section. Each section is completed by concluding comments by the editors that highlight issues covered in the section. At the end of the four sections is a discussion chapter of a set of key issues that cut-across all the chapters. These discussions among the contributing authors and the editors are targeted to addressing three key questions:

1. How do you define "generalization" and "generalizing"? What is the relationship between audiences of generalizations and the users? Who are the generalizations for? For what purpose? Are there different forms and processes of generalization? Is it possible to generalize from small scale studies?
2. What types of validity arguments are needed for generalizing in education research? Are these forms of arguments different for different forms of generalization? Can there be common ground for different generalizability arguments?
3. Given that "qualitative researchers" may count objects and members in categories and even use descriptive statistics: Do "qualitative" and "quantitative" labels serve a useful function for education researchers? Should we continue to use these labels? Do you have suggestions for alternatives, including not having dichotomous label possibilities?

The purpose of the discussion chapter is to highlight the salient issues arising from the chapters and to move our understanding to the next higher level given that each chapter constitutes a first level of learning. Taken as a whole, the introduction, overviews and highlights and the

discussion chapter constitute the main narrative of this book in which the individual arguments are embedded. This main narrative, to use an analogy, is like the body of a pendant or crown that holds together and prominently features all the diamonds and other jewels that make the piece of jewelry.

References

Bourdieu, P. (1992). The practice of reflexive sociology (The Paris workshop). In P. Bourdieu & L. J. D. Wacquant, *An invitation to reflexive sociology* (pp. 216–260). Chicago: University of Chicago Press.

Ercikan, K., & Roth, W.-M. (2006a). Constructing data. In C. Conrad & R. Serlin (Eds.), *SAGE handbook for research in education: Engaging ideas and enriching inquiry* (pp. 451–475). Thousand Oaks, CA: Sage.

Ercikan, K., & Roth, W.-M. (2006b). What good is polarizing research into qualitative and quantitative? *Educational Researcher, 35*(5), 14–23.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Roth, W.-M., & Lee, Y. J. (2007). "Vygotsky's neglected legacy": Cultural-historical activity theory. *Review of Educational Research, 77,* 186–232.

Shaffer, D. W., & Serlin, R. C. (2004). What good are statistics that don't generalize? *Educational Researcher, 33*(9), 14–25.

Section I

Generalizing Within and Beyond Populations and Contexts

Overview

Educational research relies on deconstructing[1] data about constructs such as student learning, classroom climate, and student attitudes to develop and gain insights about the education process and to inform policy and practice. Measurements such as tests, classroom observations, or interviews may facilitate this data construction effort. Most theoretical constructs education research focuses on cannot be directly observed. For example, student knowledge and skills cannot be directly observed, and what students say, do, and produce in testing situations are used to make *inferences* about these knowledge and skills. This is where much research falls short because inferences are not always supported or necessary. Thus, for example, conceptions and conceptual change researchers recognize that (a) students do not respond to instruction and (b) teachers do not take up the theory. This may not surprise some because teachers do not observe these constructs but come face to face with student talk. If we were to theorize talk-in-situation, changes therein, and teaching strategies we might obtain results that teachers can actually use.

Capturing of classroom interactions and processes through videotaping are also used to make inferences about certain target constructs such us teacher promotion of interactivity, student interest, and engagement. For data such as scores from tests to be used in a meaningful way in research, the scores need to be accurate indicators of the constructs of interest.[2] Validity of interpretations of test scores, defined as meaningfulness and appropriateness of interpretations of test scores, has played a centerpiece role in discussions of quality of research findings. Validity of interpretations of scores depends on key characteristics of tests. These include the degree to which the content covered in the test is representative of the content domain the researcher is interested in measuring. Other test characteristics are related to critical aspects of validity of interpretations. They include the questions whether tests provide consistent scores across time, raters, and test forms; whether test items

are capturing student true knowledge and skills; and whether test items depend on the format of the test items, whether they are free of culture bias. Generalizability theory (G-theory) is a statistical way of examining possible errors made in constructing the data for the research through measurement.

The most commonly understood notion of making generalization in educational research is that it denotes the making of inferences based on research in a specific context and sample to a broader set of contexts and population of subjects. Educational research defines generalizability of research findings as "external validity" (Cronbach, 1987). In other words, generalizability refers to the degree to which research claims can be extended to contexts and populations beyond those in the study itself. Even though external validity or generalizing are key components of educational research, there is not a systematic way of examining and evaluating generalizability of research findings. This problem leads to inappropriate evaluation of research generalizability based on superficial aspects of research such as sample size (small-scale versus large-scale) and methodology (statistical versus interpretive).

Generalizing and validity of inferences to a broader context are key to assessment and measurement in education. Therefore, educational measurement has systematic ways of investigating this generalizability of findings based on measurements. In particular, researchers in the area of measurement developed a statistical modeling approach to examining and estimating the degree to which inferences from test scores can be generalized beyond the testing contexts. Measurement of students' knowledge and skills are made based on a limited set of test questions and formats. Generalizability theory helps us understand to what extend scores created through measurements can be generalized beyond the set of questions and formats and to what extent the measurements represent true knowledge, abilities, and constructs that researchers are interested in. The authors assembled in this first section focus on generalization in educational research by applying the principles of G-theory to systematically think through the range of factors that may affect efforts of generalizing from educational research. The chapters in this section describe generalizability of data and how this generalizability is related to findings in research to other populations and contexts.

In their chapter entitled "Generalizability theory and its Contribution to the Discussion of the Generalizability of Research Findings" Rich Shavelson and Noreen Webb describe how Generalizability theory is related to the more general issue of generalizability of research findings. Shavelson and Webb's chapter on Generalizability theory sets us up for the subsequent chapter entitled "The Testing of English Language Learners as a Stochastic Process: Population Misspecification, Measurement Error, and Overgeneralization." In this chapter, Guillermo Solano-

Flores discusses the contribution of Generalizability theory as a theory that allows examination of sampling issues in the testing of English Language Learners (ELLs). Testing is one of the primary ways of constructing data in educational research. This research highlights the factors that might affect the inferences made based on data constructed through testing and therefore generalizability of research findings based on such data. The complexity of factors identified as relevant to the validity of test scores is a good reminder of the diversity and multiplicity of factors that affect the validity of inferences and the factors we need to consider when we examine generalizability of research findings. Previous research identified some inappropriate interpretations, generalizations, based on ELL test score data. Previous research indicates that the low test scores of ELLs often are interpreted as evidence of deficits or even disorders. For example, Richard Durán (1989) has reported that the language gap in testing has been a major contributor to the disproportionate numbers of Hispanic ELLs diagnosed as "mentally retarded" when IQ test scores were used. One study of Hispanic ELLs in Riverside, California, found that the Hispanic students, who constituted less than 10% of the school population at that time, comprised 32% of the students identified as mentally retarded (Rueda & Mercer 1985). For most of these students (62%) such decisions were based solely on low IQ scores.

Notes

1. Where, following philosophers such as Martin Heidegger and Jacques Derrida, we understand "deconstructing" to mean both taking apart (Ger. "abbauen") and preserving (Ger. "aufheben").
2. As the contributions to part C show, beyond accuracy lies the question of appropriateness and intelligibility to the target audiences.

References

Cronbach, L. J. (1987). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Durán, R. P. (1989). Testing of linguistic minorities. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 573–587). New York: American Council of Education, Macmillan.

Rueda, R., & Mercer, J. (1985). *A predictive analysis of decision-making practices with limited English proficient handicapped students*. Paper presented at the Third Annual Symposium for Bilingual Special Education, Evaluation, and Research, University of Colorado and Council for Exceptional Children, Northglenn, CO.

Chapter 2

Generalizability Theory and Its Contribution to the Discussion of the Generalizability of Research Findings

Richard J. Shavelson and Noreen M. Webb

> What's in a name? That which we call a rose By any other word would smell as sweet.
>
> From *Romeo and Juliet* (II, ii, 1–2)

What's in a name? For Romeo Montague and Juliet Capulet who meet and fall in love in Shakespeare's romantic tale … and for us … it turns out to be nothing—Romeo by any other name is Romeo … and everything—Romeo is a Montague and Juliet a Capulet, members of two relentlessly warring families. In the end, their love cannot transcend family hatred and they pay the ultimate price with their lives.

Our situation is not quite so dire. So what's in a name—oh say, "Generalizability Theory?" Nothing, it's just a name of a psychometric theory and by any other name, such as "Dependability Theory," it would be the same theory. And everything—how could a book entitled, *Generalizing from Educational Research*, not include a chapter entitled, "Generalizability Theory," regardless of the theory's content?

Now that's the question we asked ourselves when invited to contribute to this book. Our response at first was, "nothing!" The theory is not about the generalizability of research findings. Upon reflection and by analogy, however, we decided, "everything!" Well, not quite everything. Nevertheless, some of the central ideas in this arcane psychometric theory might be applied to the design of research with consequences for the generalizability of research findings. Hence we agreed to write this chapter.

Introduction

Decisions and Generalizability

When designing and carrying out empirical studies, researchers make decisions about who to study, what to study, what data to collect, and

how data will be collected and analyzed. All of these decisions have implications for the generalizability of research findings. These implications are often discussed as aspects of validity. For example, we often speak of validity of measurements, the extent to which we can generalize from scores on one measure to scores on different measures of the same or a different domain, at the current time or at a point in the future. Or we may speak of population validity, such as the extent to which research findings can be generalized from the particular sample studied to a larger (or different) population of interest. Or we may speak of ecological validity, such as the extent to which the research findings can be generalized beyond the particular environmental conditions studied to another set of conditions. Limitations arising from the researcher's decisions about the measurements to take, the population to study, and the conditions to be implemented and/or studied are often addressed in perfunctory manner in reports of findings, and may not always be recognized by researchers themselves.

Contribution of Generalizability Theory to Research Generalization

We believe that Generalizability theory, originally developed as a comprehensive approach to assessing measurement consistency—i.e., *reliability*—provides a way of making these validity issues explicit. In this paper, we show how using the lens of "G-theory" to address these validity issues can help researchers identify sources of limitations in the generalizability of their research findings (e.g., features of the measurements, the population studied, the particular instantiation of conditions in a study) and, furthermore, how G-theory provides a means of systematically investigating the extent to which these factors limit research generalization.

In a nutshell, Generalizability theory is a statistical sampling theory about the dependability or reliability of behavioral measurements. In G-theory, a person's score on a measurement (e.g., science test) is considered to be a sample from an indefinitely large universe of scores that person might have earned on combinations of other test forms, on other occasions, scored by other raters. Reliability, then, is an index of just how consistently we can generalize from the sample of measurements in hand to the universe of interest. That is, it is an index of how accurate the inference is *from* a person's score on this particular form of the science test given on this particular occasion as scored by this particular rater *to* this person's average score earned if she had taken all possible test forms on all possible occasions scored by all possible raters. G-theory, then, views reliability as the accuracy with which we can generalize from a

sample (a single test score) to the universe of interest defined by the average score over all possible forms, occasions and scorers.

It seems to us that the question of generalizing education research findings from a sample of measurements in hand to a larger universe of interest can, at least in part and for some purposes, be conceived in a similar way. How well can we generalize from the sample of measurements in hand to a broader domain? We believe that the kind of reasoning that underlies G-theory would at least be heuristically useful in thinking about the design of research for certain kinds of inferences, whether we are speaking about large statistical samples or small case studies.

In what follows, we sketch G-theory in a bit more conceptual detail, leaving aside completely the statistical developments (see G-theory references above). We believe that notions underlying G-theory such as the "universe of admissible observations," "universe of generalization," "random and fixed facets," and "crossed and nested designs" have much to say about the design of research. Once we have laid out these fundamental notions, we then draw parallels from G-theory to generalization in education research.

Some Fundamental Ideas from Generalizability Theory

In G-theory a behavioral measurement (e.g., a test score) is conceived as a sample from a *universe of admissible observations*. This universe consists of all possible observations that decision makers consider to be acceptable substitutes (e.g., scores sampled on occasions 2 and 3) for the sample observation in hand (scores on occasion 1). A measurement situation, then, can be characterized by a set of features such as test form, test item, rater, or test occasion. Each characteristic feature is called a *facet* of a measurement. A universe of admissible observations, then, is defined by all possible combinations of the levels of the facets (e.g., all possible items combined with all possible occasions).

Generalizability Study

A generalizability (G) study is like a "pilot study" that is designed to isolate and estimate as many facets of measurement error in the universe of admissible observations as is reasonably and economically feasible. The study includes the most important facets that a variety of decision makers might wish to generalize over (e.g., items, forms, occasions, raters). This explicit, full formulation of the universe, some or all of which a particular decision maker might generalize to, might prove useful to researchers concerned about research generalization. In some senses, making explicit the universe of admissible observations provides

a vision of what ultimately research such as the particular study in hand is intended to tell its audiences about.

To be concrete, suppose that, in studying science achievement the universe of admissible observations is defined by all possible combinations of items, raters and test occasions that a variety of decision makers would be equally willing to interpret as bearing on students' science achievement. Ideally, a G-study would include all three facets (item, rater, test occasion). For example, a random sample of students would be tested on 3 test items randomly sampled from a large domain of such items and 3 randomly sampled raters would score their performance on 2 randomly selected occasions (see Table 2.1). Depending on which multiple-choice alternative was selected the student could earn an item score ranging from 1 to 5 points. The test was administered twice over roughly a two-week interval.

In this G-study, student (person) is the *object of measurement*[1] and both item and occasion are *facets* of the measurement.[2] The test items and occasions in the G-study constitute a sample from all possible items and occasions that a decision maker would be equally willing to interpret as bearing on students' science achievement. To draw a parallel to research generalization, note that: (a) the object of measurement corresponds to the population to which a researcher wishes to generalize, (b) the facets correspond to the treatment conditions and (say) organizational contexts to which she wishes to generalize, and (c) the item sample corresponds to the universe of science content, knowledge and skills to which the researcher wishes to generalize.

To pinpoint different sources of measurement error, G-theory estimates the variation in scores due to each person, each facet, and their combinations (interactions). More specifically, G-theory estimates the *components of observed-score variance* contributed by the object of

Table 2.1 Person × Item × Occasion G-Study of Science Achievement Scores

Person	*Item*	*Occasion*					
		I			*II*		
		1	*2*	*3*	*1*	*2*	*3*
1		3	1	5	4	3	4
2		4	1	4	4	2	3
3		2	3	3	3	2	4
…							
p		4	5	4	4	4	2
…							
n		2	4	4	3	4	3

measurement, the facets, and their combinations. In this way, the theory isolates different sources of score variation in measurements. In a similar manner, research generalization might attend to estimating the "effects" of person, treatment and content sampling.

To be concrete about estimating effects, continuing with the science test example, note that the student is the object of measurement and each student's observed score can be decomposed into a component for student, item, occasion, and combinations (interactions) of student, item, and occasion. The student component of the score reflects systematic variation in their academic ability, giving rise to systematic variability among students (reflected by the student or person variance component). The other score components reflect sources of measurement error. For example, a good occasion (e.g., following a school-wide announcement that the student body had received a community award for reducing environmental hazards based on their science experiments) might tend to raise all students' achievement, giving rise to mean differences from one occasion to the next (indexed by the occasion variance component). And the particular wording of an item might lead certain students to answer incorrectly compared to other students, giving rise to a non-zero person x item interaction (p x i variance component).

Decision Study

The Decision (D) study uses information from the pilot study—the G-study—to design a measurement procedure that minimizes error for a particular purpose. In planning a D-study the decision maker defines the *universe of generalization*, which contains the facets (and levels of them) over which the decision maker proposes to generalize. A decision maker may propose to generalize over the same facets (and levels of them) as in the universe of admissible observations (e.g., item, occasion, rater). Another decision maker, however, may propose to generalize less broadly than the universe of admissible observations because of time, cost, or particular interest (e.g., a decision maker is only interested in students' spring science achievement). That is, a decision maker may propose to generalize over only a portion of the universe of admissible observations. In this case, the universe of generalization is a subset of the universe of admissible observations—the set of facets and their levels (e.g., items and occasions) to which the particular decision maker proposes to generalize.

What the particular decision maker would ultimately like to know about a student is his or her *universe score*—defined as the long-run average of that student's observed scores over all observations in the decision maker's universe of generalization. The theory describes the dependability ("reliability") of generalizations made from a person's

observed score on a test to the score he or she would obtain in this universe of generalization—to his or her "*universe score*." Hence the name, "Generalizability Theory."

A decision maker's universe of generalization (and hence the design of the D-study) may be narrower than the universe of admissible observations for a variety of reasons. Consider a universe of generalization restricted to one facet, say, items. In this case, multiple items would be used but only one test occasion (e.g., the spring test administration) would be used in the D-study and generalization would not be made from the spring test scores to scores that might have been obtained on another occasion. Some decision makers may choose to hold constant occasion (spring testing) because they would like to know how many items are needed on the science achievement test to produce a trustworthy sample of a student's spring science achievement. Other decision makers may be interested in generalizing over occasions but decide to restrict attention to one test occasion because it turns out to be too expensive or time-consuming to obtain scores from multiple occasions in the D-study. Or the G-study may show that you would need to have too many occasions for decent generalizability and the decision maker throws up his arms and says, "Forget about generalizing across occasions!"

From a research generalization perspective, G-theory's lesson might be its insistence on clarity between the universe of admissible observations—perhaps a comprehensive ideal—and the practical reality of resource constraints—the universe of generalization. Being explicit about the differences between these two universes might make clear to researchers and more general audiences the extent and limits of research generalization.

Generalizability and Decision-Study Designs

Generalizability theory allows the decision maker to use different designs in G- and D-studies because the two types of studies have different goals. G-studies attempt to estimate as many variance components as possible in the universe of admissible observations so as to be useful to decision makers with different goals. D-studies attempt to meet decision makers' goals while economizing on facets to get the biggest bang (reliability!) for a constrained buck. Again, this explicit representation of universes might prove useful to research generalization.

Designs with Crossed and Nested Facets

Typically in a G-study, a *crossed* design is used. In a crossed design, all students are observed under each level of each facet. This means that, in our example, each student responds to each science-test item on each

occasion. The crossed design provides maximal information about the components of variation in observed science scores. In our example, seven different variance components can be estimated—one each for the main effects of person, item, and occasion; two-way interactions between person and item, person and occasion, and item and occasion; and a residual due to the person x item x occasion interaction and random error.

Information from the G-study, in the form of variance components, can be used to design a D-study by projecting the impact of changing the number of levels of a facet on the reliability of the measurement. Consider the case of the decision maker choosing to hold constant occasion (spring testing) and seeking to know how many items are needed on the science achievement test to produce a trustworthy sample of a student's spring science achievement. It is well known that the sampling variability of test items, especially in interaction with the object of measurement, person (p x i), is very large. By focusing on the facet, item, the decision maker can determine the number of test items needed to reach some level of reliability, say 0.80 (on a scale ranging from 0 to 1.00).

When more than one facet is a major source of measurement error the decision maker might want to project the tradeoff in reliability by varying the number of items *and* the number of occasions. When more than one facet is considered in a D-study—e.g., the decision maker is interested in students' science achievement any time from March to June—information from the G-study can be used to evaluate the tradeoff of increasing items on the test or the number of test occasions.

Finally, the decision maker might be concerned with the amount of testing time, especially if the test were to be given on two occasions. In this case, she might consider testing a different subset of science items on each of two different occasions. In this last example, and as is common in D-studies, we say that test items (subtests) are *nested* within occasions—items 1 to 20 are administered at occasion 1 and items 21 to 40 are administered at occasion 2. With this nested design, the decision maker can hold total testing time constant, while administering a broader array of items (40 items) than if the same 20 items were administered on both occasions (a *crossed* design).

While G-studies typically employ or should to the extent feasible employ crossed designs in order to estimate each and every possible source of variation in a student's test score, D-studies may profit by using nested designs which are economical and efficient and can be used to increase the levels of a particularly cantankerous facet[3] within reasonable cost constraints. The parallel to research generalization, it seems to us, is for researchers to decide on which of a variety of designs would meet requirements for inferring to their universe of generalization while maximizing the power of their statistical tests.

Designs with Random and Fixed Facets

G-theory is essentially a random effects theory—inferences are drawn from a random sample in hand to what is the case in an indefinitely large, carefully defined universe of possible observations. Typically, a random facet is created by randomly sampling levels of a facet.[4]

A fixed facet (cf. fixed factor in analysis of variance) arises when the decision maker: (a) purposely selects certain levels of the facet and is not interested in generalizing beyond them, (b) finds it unreasonable to generalize beyond the levels observed, or (c) when the entire universe of levels is small and all levels are included in the measurement design. A fixed facet, then, restricts the decision maker's *universe of generalization*. G-theory typically treats fixed facets by averaging over the levels of the fixed facet and examining the generalizability of the average over the random facets. When it does not make conceptual sense to average over the levels of a fixed facet, a separate G-study may be conducted within each level of the fixed facet.

To see how fixing a facet might work, consider a study of teaching behavior in which elementary teachers are observed teaching mathematics and reading. These are two subjects in a potentially broad array of subjects that might be taught in elementary school. Because we believed that teaching behavior in mathematics and reading may not be generalizable to teaching behavior in other subjects, we considered "subject" to be a fixed facet in the *universe of generalization*. Moreover, we reasoned that teaching mathematics is considerably different from teaching reading. As a consequence, we conducted separate D-studies for mathematics and reading scores.

Designs and the Object of Measurement

The discussion up to this point has treated person as the object of measurement. However, the focus of measurement may change depending on a particular decision maker's purpose, as described in the *principle of symmetry:* "The principle of symmetry of the data is simply an affirmation that each of the facets of a factorial design can be selected as an object of study, and that the operations defined for one facet can be transposed in the study of another facet" (Cardinet, Tourneur, Allal, 1981, p. 184). In a persons (*p*) x items (*i*) x occasions (*o*) design, whereas persons may be the focus of measurement for evaluators wishing to make dependable judgments about persons' performance, items may be the focus for curriculum developers wishing to calibrate items for use in item banks. In the latter case, individual differences among persons represent error variation, rather than universe-score variation, in the measurement.

Moreover, the object of measurement may be multifaceted, for example, when educational evaluators are interested in scholastic achievement of classes, schools, and districts, or in comparisons across years. Or the focus may be on items corresponding to different content units in which the universe-score of interest is that of items (*i*) nested within content units (*c*). Or objects of measurement may be defined according to attributes of persons, such as persons nested within geographic region, gender, or socio-economic status. We treat the concept of multifaceted populations more fully below.

From a research generalization perspective, we see two related lessons from G-theory. The first lesson is that generalization depends on how the object of measurement is sampled, how treatment and context are sampled, and how the outcome measurements are sampled. We see great attention paid to sampling the object of measurement but little attention paid to sampling treatments/contexts or measurements in substantive research. Yet such sampling has a great deal to say about the credibility of generalization from treatments/contexts or achievement tests in hand and the broader universe of generalization. The second related lesson is that, as we will show below, treatments/ contexts are often implicitly sampled (not fixed) and yet such sampling is not taken into account in research generalization.

Generalizability Theory and Validity

G-theory has focused on estimating multifaceted measurement error and reliability. It extended traditional reliability theory by going beyond separate estimates of reliability: internal consistency ("Cronbach's alpha"), test-retest, and alternate or parallel forms, inter-rater reliability. G-theory includes as *typical facets* of a measurement within in an overarching framework—items (alpha), occasions (test-retest), forms (alternative forms) and raters (inter-rater). The theory then statistically estimates the contribution of these facets individually and in combination with each other and the object of measurement—to estimate the (in)consistency of measurement simultaneously.

If we move beyond the typical measurement facets associated with reliability of test (and other) scores—item, form, rater, occasion—to include facets such as type of test (multiple-choice, short answer, performance assessment), we have moved outside the traditional boundaries of reliability and generalizability theory into areas of validity. In particular, in this case, we have moved to convergent validity, asking, "To what extent do different measurement procedures purported to measure the same attribute converge and give the same 'picture'?" Or, "To what extent do the sample of test items generalize to the broad content domain?" an