The final version of this article was published as: Ercikan, K., & Roth, W.-M. (2014). Limits of generalizing in education research: Why criteria for research generalization should include population heterogeneity and users of knowledge claims. *Teachers College Record, 116*(5), 1–28.

Limits of Generalizing in Education Research: Why Criteria for Research Generalization Should Include Population Heterogeneity and Uses of Knowledge Claims

Kadriye Ercikan, *University of British Columbia*

Wolff-Michael Roth, *University of Victoria*

Abstract

Generalization is a critical concept in all research that describes the process of developing general knowledge that applies to all elements of a unit (population), while studying only a subset of these elements (sample). Commonly applied criteria for generalizing that focus on experimental design or representativeness of samples of the population of units neglect considering the targeted uses of knowledge generated from the generalization. This paper (a) articulates the structure and discusses limitations of different forms of generalizations across the spectrum of quantitative and qualitative research; and it (b) argues for an overarching framework that includes population heterogeneity and uses of knowledge claims as part of the rationale for generalizations from educational research.

A recent special issue of this journal was dedicated to data use as an integral part of current reform efforts (Turner & Coburn, 2012). Other researchers highlight data

[v] Accepted for publication in *Teachers College Record*.

use and research evidence as perhaps the most central dimension of today's political climate that shapes the field of education (Cohen-Vogel, 2011; Moss, 2012; Roderick, 2012). This emphasis on data use and evidence crowns empirical research findings with the highest status in guiding policy and practice. It is therefore legitimate to ask, "To what extent is typical education research designed to provide evidence to inform policy and practice?" The evidence educators use for educational policy analysis, evaluation, and decision-making tends to be produced through educational research that takes population samples or case studies to make claims valid for jurisdictions at different levels such as classroom, schools, districts etc. However, the question whether research evidence at one level of educational practice scales up to another level is a non-trivial question (e.g., Ercikan & Roth, in press; Stein et al., 2008). The question of the extent to which educators can generalize from educational research has led in many contexts to a predilection for experimental and quantitative over qualitative studies – although it has been suggested that without the examination of qualitative evidence, "variations in quantitative studies are difficult to interpret" (Kennedy, 2008, p. 344). But in education and other fields, "[f]indings from a large number of qualitative research investigations have had little impact on clinical practice and policy formation" (Finfgeld-Connett, 2010, p. 246). In this article, we argue that the issue of generalization of empirical findings for the purpose of education practice, policy analysis, evaluation, and decision-making not only needs to transcend the traditional divide between quantitative and qualitative research but also requires an overarching framework that includes population heterogeneity and uses of knowledge claims as criteria that establish the quality of generalizations that meets policy makers' demands "for relevant and rigorous research" (Brewer, Fuller, & Loeb, 2010, p. 4). In so doing, we contribute to establishing a theoretical framework for methodological rigor related to educational research generalization.

Introduction

The level and kind of evidence education research produces has been at the forefront of educational reform for several decades now; in particular, it had been a central issue in George W. Bush's education reform agenda with *No Child Left Behind* and more recently with Barack Obama's *Race to the Top* reform efforts. In the wake of these reform agendas, the Institute for Education Sciences was created "to advance the field of education research, making it more rigorous in support of *evidence-based* education and therefore a critical component to the success of No Child Left Behind" (USDE, 2004, our emphasis). Although the term "evidence-based" often has been taken to mean the result of experimental and large-scale statistical studies, the question "What constitutes evidence?" is much more complex than that. What type of evidence is needed to support decisions about effective programs or actions that can help improve learning? Can the same evidence support such decisions for different groups and individuals? To answer these questions, we need to consider the extent to which educational research conducted in one setting generalizes to other settings, from a sample to the target population, to a sub-population, and to individuals. In a policy context that places great value on *evidence-based* research, experimental studies and investigations using high-power statistics tend to be privileged as having the capacity to support generalizations that can contribute to sound decision making and planning (Maxwell, 2004; Slavin, 2008; Song & Herman, 2010). This possibility to generalize from research findings is a primary factor in determining the value and importance of research (e.g., Ercikan & Roth, 2009). [1] However, limited notions of generalization from educational research – which tend to be tied to representativeness and size of the sample of students, schools, and settings used in research – lead to mistakes about large-

sample studies being more generalizable than research conducted using small sample sizes (typically qualitative research, ethnographic studies) and overlook limitations of these generalizations for informing policy and practice.

We define generalization as development of knowledge claims from education research based on a limited set of settings, contexts, conditions, and samples. Even though in all generalization the process involves making knowledge claims based on specific to general, different rationales and criteria for generalization claims may be used. Research generalization is typically considered as part of external validity (Campbell, 1986; Cronbach, 1982). The outcome of the generalization process is knowledge that may describe, explain, and guide educational processes in contexts other than those investigated in a specific research project. But depending on the uses, the targeted *levels* of the required knowledge will be different. When uses of such knowledge is informing policy, the targeted level of knowledge is the group – such as fourth grade students in the country, English Language Learners (ELLs), or special needs students (Bachman, 2009). On the other hand, when the use of such knowledge is to inform teachers, practitioners, or parents who are working with individual students, the targeted knowledge level is the individual student. Consequently, different levels of knowledge claims and therefore evidence about effective strategies are needed for research to inform policy and practice in meaningful ways (Luke, 2009). Group level knowledge such as whether an educational intervention is effective based on an experimental design likely is insufficient for making decisions about effectiveness of the intervention for individual students or for sub-groups of students such as males or females, ELL, or students at different ability and achievement levels. This is so not only because of the statistical nature of experimental design but also because almost all experiments are based on the logic of inter-individual differences and co-variations rather than

on a logic of within-individual differences and causations (Borsboom, Mellenberg, & van Heerden, 2003).

This article has two connected purposes: (a) to articulate the structure and discuss limitations of different forms of generalizations across the spectrum of quantitative and qualitative research and (b) to argue for considering population heterogeneity and for including future uses of knowledge claims when judging the appropriateness of generalizations that are used as evidence on which educational policy analysis, evaluation, and decision-making are based. In the first part of the paper we present two forms of generalization that rely on statistical analysis of between-group variation: analytic and probabilistic generalization. These are the most commonly understood notions of generalizing in educational research (Eisenhart, 2009; Firestone, 1993). We then describe a third form of generalization: essentialist generalization. [2] Essentialist generalization moves from the particular to the general in small sample studies. This form of generalization exists in medical, (historical-) genetic, and scientific research in general, but is not well understood and is infrequently used in social science or education research. We discuss limitations of each kind of generalization and propose two additional criteria when evaluating the validity of evidence based on generalizations from education research. In the second part of the paper, we first make a case for taking into account population heterogeneity when evaluating validity of generalizations from educational research. Second, we demonstrate a need to consider future use as integral and essential aspects of the question about the extent to which research claims are generalizable.

Generalizing in Educational Research

In this section we present and discuss – cutting across the quantitative-qualitative divide that exists in educational research methodology – three main forms of generalization and their limitations in view of how they inform different users in policy and practice. The three **forms of generalization** *analytic*, *probabilistic,* and *essentialist* are presented as distinctly different with respect to the rationale and evidence required to support them. The **criteria** used for judging the supporting evidence are described. The distinctions between the three forms of generalization are important to clarify in discussing limitations of each generalization in informing policy and practice. None of them are presented as superior to the other; rather they are considered as complementary.

Analytic Generalization

Structure*.* Analytic generalization relies on the design of the research to make causal claims. It involves making arguments that support claims in relation to a theory. It may involve the testing of a new theory as well as application of a theory in a context for which the theory was not originally developed. The researcher may hypothesize, for example, that an intervention operationally defining a theoretical construct leads to better learning. This operationalization requires a specific research design (Shadish, Cook, & Campbell, 2002). First, it must logically allow making *causal* inferences: Instances where a cause operates have to lead to significantly different observations than those instances where the cause is disabled. Usually, this requires randomly assigning participants to control and experimental groups in the hope of achieving equivalence of these groups with respect to all moderating and mediating variables and an identical implementation of the intervention to the experimental and comparison (control) groups. The groups are not expected to be representative samples of any particular target population.

Random equivalence is intended to rule out any potential alternative explanations of differences between the control and experimental groups. The arguments in analytic generalization are closely tied to the degree to which experimental design is truly implemented. The statistical support for the hypothesis about the effectiveness of the interventions – which provides sufficient evidence to reject the null hypothesis that there is no difference between the control and the experimental groups after the intervention – is used to make claims about effects of the intervention in the target population. The claim is with respect to the causal relationship between the intervention and the outcome. The outcome of an intervention is determined by comparing the difference between the means of control and experimental groups to the standard error of the mean differences. If on the average a statistically significant difference in the hypothesized direction is identified between the two groups, the theory is supported and therefore implies effectiveness of the intervention, such as a new instruction method that includes using technology in mathematics teaching.

Limitation. In analytic generalization, there are **two key criteria** for judging the causal inference from the experimental design. One is whether there is a systematic difference between experimental and control groups that can be supported by statistical evidence and the other is the degree to which a true experiment has been conducted so that the change in experimental group outcomes can be attributed to the specific operating cause deriving from the intervention. Even when such a generalization is fully supported based on these two criteria, a loose causal link is established. A causal claim that applies to the overall group does not necessarily apply to subgroups or to individuals because the logic of such studies is based on the logic of between-subjects rather than within-subjects variation (Borsboom et al., 2003). In other words, intervention may have been effective “on the average” but,

because the theory and measurement models are not based on within-individual causation, the latter may not apply to some individuals and not for some sub-groups. Figure 1 presents distributions of outcome scores for experimental and control groups from a hypothetical experiment. As the overlapping area in Figure 1 shows, a considerable number of individuals in the control group may perform higher than individuals in the experimental group (overlapping area). Even though individuals from the experimental group are more likely to be on the higher end of the scale and those from the control group are more likely to be at the lower end of the scale, we cannot tell how the change in scores varied for different individuals or sub-groups and whether the change was uniformly in the same direction. The degree of change and the direction of change for individuals in the experimental group cannot be determined by comparing score distributions with the control group.

«««««« Insert Figure 1 about here »»»»»»

Research on aptitude-treatment interactions (e.g., Corno et al., 2002; Snow, 1989) shows that claims based on group score distributions do not tell us whether this program was effective for certain groups, for example ELL students, males or females, or for individual students. For example, one design experiment [3] investigating a science-through-artifact-design approach showed that most of the learning disabled students ended up scoring in the top quartile, whereas those students who did well in traditionally conducted science classes ended up scoring much lower on the different tests used to test their understanding (Roth, McGinn, Woszczyna, & Boutonné, 1999). This exemplifies that some individuals may not have benefited from a well-intended intervention and some others may have been hurt or negatively affected by the intervention because of treatment by aptitude

interactions. Thus, although the treatment has led to an overall effect, it cannot be concluded that the treatment should be used with and is applicable to any subgroup or individual. Even though sub-group analyses by gender or language groups may reveal whether broad claims about relations apply to these sub-groups, often, the sample size for the subgroups is not sufficient to adequately address the effectiveness of the intervention by sub-groups. Furthermore, variations across groups may not be limited to easily identifiable demographic groups. The levels of effectiveness of programs for sub-groups are rarely considered as part of policy decisions. In summary, even though the only research design that allows making causal claims is commonly accepted to be experimental design, such designs in fact do not provide evidence of a causal link for sub-groups or individuals. Therefore, the research results cannot inform practitioners who are dealing with individuals or policy makers who deal with specific sub-groups for whom the research results are not explicitly identified regarding the effectiveness of treatments and interventions.

Probabilistic (Sample to Population) Generalization

Structure. Probabilistic generalization – also known as statistical or sample-to-population generalization – relies on representativeness of a sample of a target population. It is used to describe population characteristics and does not include causal claims (Eisenhart, 2009; Yin, 2008). Researchers and consumers of research judge knowledge claims by the degree to which samples of subjects, outcomes, and contexts used in research are representative of the populations to which the research is intended to generalize (Ercikan, 2009; Firestone, 1993). The logic of this form of inference is an ideal type of induction, which moves from the concrete observation (a feature of the sample) and, via inducing the case (a feature of the *entire* population), arrives at the general knowledge claim. Two broad types of

probabilistic generalizations are common. One type of generalization claim is with respect to relationships between variables, for example, between IQ and achievement (Figure 2a). In this case, statistics is used to estimate the probability that a systematic relation between IQ and achievement exists beyond chance level. The second type of research generalization is related to relative frequency (e.g., proportion of students identified with learning disabilities) or group differences (e.g., differences in achievement between boys and girls) (Figure 2b). For example, the Programme for International Student Assessment (PISA) 2009 data for Canada suggest that there are statistically significant differences between boys and girls on the reading score (M_b = 507, M_g = 542, $SD_{b,g}$ = 90) (see Figure 2b) based on the differences in the sample. In both of these probabilistic generalizations, generalization claims are derived from observations from the sample. The **criteria** by which the generalization is judged – i.e., the validity of claims about the correlation between IQ and achievement or gender differences in reading in Canada – centers on one of the same criteria used for judging analytic generalization that is whether there is statistical evidence of a systematic pattern in the data. Even though probabilistic generalizations may include group comparison, such as comparing gender or ethnic groups, these generalizations do not require a specific research design such as random equivalence of groups, or standardized implementation of an intervention. Instead, the representativeness of the samples of the target populations is the second key criterion used for probabilistic generalizations.

Limitation. Within group heterogeneity that limits the meaningfulness of causal claims in analytic generalization for sub-groups or individuals leads to similar limitations in probabilistic generalization. National surveys of achievement are primary data sources for making probabilistic generalizations. For example, one of the primary foci of large-scale surveys of achievement – e.g., the National

Assessment of Educational Progress (NAEP) or international assessments such as PISA – is to compare outcome levels of males and females, countries, or ethnic groups. Using the recent PISA reading results, we plotted the distribution of reading scores for Canadian boys and girls (Figure 2b). These distributions of scores have a great degree of overlap, so that claims such as "girls are outperforming boys" are not meaningful. At each score level, we find boys and girls, though at higher scoring levels, there are more girls than boys with a given score (right, Figure 2b), whereas there are more boys than girls with a given score at lower scoring levels (left, Figure 2b). Which girls are outperforming which boys? Clearly some boys are outperforming some girls. In fact, as recent results in the UK show, although girls tend to exhibit higher achievement scores on average (e.g., number of A's in A-level courses), there are more boys than girls among the very highest scoring students (Clark & Preece, 2012). Thus, the claims for generalizing group differences become even more complex and problematic when we look at gender differences between sub-groups such as those from different socio-economic background, language groups, and others. A similar limitation exists when making knowledge claims related to relationships between variables. Probabilistic generalization that focuses on describing population characteristics can lead to knowledge claims that involve statistical concepts – e.g., mean, frequency, mean differences, or correlations – may not apply to sub-groups and may have limited value for guiding policy and practice.

«««««« Insert Figure 2 about here »»»»»»

Essentialist Generalization

Structure. Essentialist generalization is the result of a systematic interrogation of "the particular case by constituting it as a 'particular instance of the possible' . . .

in order to extract general or invariant properties that can be uncovered only by such interrogation" (Bourdieu, 1992, p. 233). In this approach, *every* case is taken as expressing the underlying law or laws; the approach intends to identify invariants in phenomena that on the surface look like they have little or nothing in common (Roth, 2012). Thus, for example, Vygotsky (1971) derived a general theory of the psychology of art based on the analysis of three very different literary genres: a fable, a short story, and a tragedy. He concludes:

> We have ascertained that contradiction is the essential feature of artistic form and material. We have also found that the essential part of aesthetic response is the manifestations of the affective contradiction which we have designated by the term *catharsis*. (p. 217, original emphasis, underline added)

Having derived his psychology of art based on individual case studies generally and the role of catharsis more specifically, Vygotsky notes that "it would be very important to show how catharsis is achieved in different art forms, what its chief characteristics are, and what auxiliary processes and mechanisms are involved in" (p. 217). That is, although Vygotsky developed the categories of *affective contradiction* and *catharsis* and their role in human development from the analysis of a concrete case, which he subsequently verifies by means of analogy in two further cases, he arrives at generalizations that are much broader than the three texts he analyzed and much broader than the written forms of art. Thus, as shown in Figure 3, because the categories constitute the essential feature of artistic form and material they equally can be found in painting and music (blues, classical, or any other form). In a subsequent text he summarily states: "the principle of art as well is dealing with a reaction which in reality *never manifested itself* in a pure form, but

always with its 'coefficient of specification'" (Vygotsky, 1927/1997, p. 319). This is so because he has abstracted, for example, *from* the concrete characteristics of the fable to derive at "the essence of the aesthetic reaction" (p. 319). To know what the generalization means in any particular situation, therefore, requires finding "the factual boundaries, levels and forms of the applicability" (p. 319); and this "is a matter of factual research" (p. 319). Vygotsky suggests that this is the task of historical research, as it can show "*which* feelings in *which* eras, via *which* forms have been expressed in art" (p. 319, original emphasis). In this approach it is crucial, therefore, not to universalize the particular case but to reveal the invariant properties that hide themselves under the appearance of the singularity of each case (see also Bourdieu, 1992; Mannheim, 2004). The invariant properties derive from the fact that there is a *common history* underlying the different cases. Essentialist generalization tends to identify the work and processes that produce phenomena rather than the phenomena themselves (Roth, 2012). Thus, although queues manifest themselves in a multitude of ways – at a supermarket checkout counter, movie theater guichet, freeway on-ramp, bus stop, or passport office with ticket system – once the structure of the queuers' *work* (i.e., method) has been identified, every single case of queuing whatever its particular context is understood (Garfinkel, 2002).

This type of arriving at explanations and the form of generalization often takes a historical-developmental perspective whereby the "general" (common) is an evolutionary earlier form (Il'enkov, 1982). Marx [Engels] (1962) developed this form of thinking using the concept "man"; Leontyev (1981) and Holzkamp (1983) used this method in a categorical reconstruction of the human psyche. As the common concretizes itself in subsequent generations, it diversifies, expressing the possibilities that exist in the general. Research, moving in the opposite way of history, posits a possible generalization based on the observation (i.e., rule, law) *and*

then moves by way of deduction to the case and observation that the generalization implies. In the case of a divergence between actual observation and the one deduced from the posited generalization, the latter is modified and tested again. Essentialist generalization is intended to produce meanings that pertain to other fields of observation. The **criterion** by which it is judged is *testability,* which is designed to guarantee invariant nature of the categories derived (Bourdieu, 1992; Vygotsky, 1927/1997). Here, *testability* refers to the fact that a tentative statement of generalization can be tested by examining any other concrete case. Thus, the initially tentative laws that Vygotsky (1971) identified in the fable can be tested in any other art form; with such tests, researchers are "verifying the formula" (p. 217).

«««««« Insert Figure 3 about here »»»»»»

Limitation. In the debate about what constitutes evidence, experimental and quantitative investigations are often preferred because, so goes the charge, qualitative research does not generalize. [4] The problem does not lie with qualitative research as such. The real problem is two-fold: (a) some qualitative researchers do not attempt to find invariants for understanding human behavior across different settings while (b) others overgeneralize by universalizing the particulars of the field of observation to other situations (Bourdieu, 1992). Moreover, the limitations of this approach do not lie with the method, because, in contrast to its analytical and probabilistic cousins, essentialist generalization is intended to yield claims that apply to *every* case. As with any other form of generalization, essentialist generalization leads to statements that require specification in particular contexts. To achieve generalization in case-based research, the particular has to be taken *as* a particular and it has to be generalized "to discover, through the application of general questions, the invariant properties

that it conceals under the appearance of singularity" (p. 234). One achieves this by completely immersing oneself "in the particularity of the case at hand without drowning in it . . . to realize the intention of *generalization* . . . through this particular manner of thinking the particular case which consists of actually thinking it as such" (pp. 233–234). Case-based research too frequently does not lead to generalization and furthermore "inclines us toward a sort of *structural conservatism* leading to the reproduction of scholarly doxa" (p. 248). In the case of phenomenography, researchers tend to catalogue the kinds of experiences research participants have but tend to fail seeking generalizations that would *explain why* participants experience a situation in this or that manner under given conditions (e.g., Roth, 2009a, 2009b).

Additional Research Generalization Criteria: Population Heterogeneity and Uses

> [I]n the case of stating truly or falsely, just as much as in the case of advising well or badly, the *intents* and *purposes* of the utterance and its *context* are important; what is judged true in a school book may not be so judged in a work of historical research. (Austin, 1962/1975, p. 143, emphasis added)

The criteria for generalization – i.e., the types of evidence needed to support knowledge claims – vary in different types of generalizations. In analytic generalization, the key criteria are (1) whether a systematic difference between experimental and control groups can be supported by statistical evidence and (2) whether the change in experimental group outcomes can be causally linked to the intervention. In probabilistic generalization, the key criteria are (1) whether systematic patterns in the sample can be supported by statistical evidence and (2)

whether the sample is representative of the population. In essentialist generalization, the degree to which essential (i.e., common to all cases) aspects of the case are found in other cases of people, interventions, and contexts determine whether generalization claims are supported. To what extent are these currently used criteria for research generalization sufficient for determining meaningfulness and applicability of knowledge to inform policy and practice?

In analytic generalization, the causal claim "the intervention causes the difference between the control and experimental groups" or in the probabilistic generalization "girls are performing higher than boys in the reading assessment" are targeted to be at the group level. Generalization of such claims is based on statistical analysis of between-group variation – also referred to as the "variable model" (Holzkamp, 1983; Maxwell, 2004) or the "snapshot, bookend, between-groups paradigm" (Winne & Nesbitt, 2010, p. 653). This approach entails within-group homogeneity. Researchers have criticized the use of between-group analyses for making claims about within-individual processes. Thus,

> there is an almost universal – but surprisingly silent – reliance on what may be called a uniformity-of-nature assumption in doing between-subject-analyses; the relation between mechanisms that operate at the level of the individual and models that explain variation between individuals is often taken for granted, rather than investigated. (Borsboom et al., 2003, p. 215)

A great deal of other research findings parallel this position (cf., Ercikan, Roth, Simon, Sandilands and Lyons-Thomas, in press; Molenaar, 1999, 2004; Molenaar, Huizenga, & Nesselroade, 2003; Oliveri, Ercikan & Zumbo, in press a; Oliveri, Ercikan & Zumbo; in pressb). These findings demonstrate that "if a model fits in a given population, this does not entail the fit of the same model for any given element

from a population, or even for the majority of elements from that population" (Borsboom et al., 2003, p. 213). Similarly, qualitative research often fails to recognize that in the apparent diversity of phenomena there are fundamental commonalities in the processes of their generation (Garfinkel, 2002; Vygotsky, 1927/1997).

In our introductory quotation to this section, Austin points out that to establish the truth or falsity of a statement we need to know its context intents and its and purposes (i.e., uses). In this section we introduce two additional criteria when making generalization claims that address context and use of knowledge claims. Respectively, these are (a) heterogeneity in the target population and (b) the degree to which claims apply to the targeted uses.

Population Heterogeneity as a Criterion in Research Generalization

The question about the degree to which some research claim provides useful direction for practice and policy depends on the degree to which findings apply to the relevant sub-groups or individuals. Applicability of research findings to the relevant units (individual, sub-group, group) is at the core of potential for research to inform pedagogy, policy, or social theory. Research inferences targeted to broadly defined populations have significant limitations in their applicability to understanding or to making decisions regarding sub-groups of the populations such as gender, ethnic, and ability groups of students. Cronbach (1982) highlights diversity in the population and its potential effect on inferences by stating that "the summary statistics on the sample, or the estimates for UTOS or a sub-UTOS, are usually not an adequate base for inference about *UTOS. Insofar as there is diversity in the data, the consumer should be told about that diversity and any factors associated with it" (p. 167). [5] As a result, the researcher will have "to work back

and forth between the gross statistical analysis and the differentiated, select cases, taking one level of analysis as background for the other" (p. 167). This approach, requires an explicit recognition of heterogeneity in the population and an examination of the degree to which sub-group results deviate from the overall group results.

Educational theorists and researchers interested in making generalizations always face a dilemma between the general and the particular because "fine partitioning allows more accurate predictions; and broad categories, less accurate but widely applicable generalization" (Corno et al., 2002, p. 227). This is the situation even when a generalization is true *in every case*, because the general in the form of type, norm, or limit "will never manifest itself in exactly that form. But the type, norm, or limit will always be part of the concrete reaction and determine its specific character" (Vygotsky, 1927/1997, p. 319). [6] However, most widely used research methods that rely on between-group variation do not take this variation into account when a generalization is made (Molenaar, 2004). In particular, this is relevant to several types of research that focus on education improvement. Typical educational intervention models focus on change identified at the level of group outcomes and attempts to evaluate the effectiveness of the intervention using experimental or quasi-experimental designs. The intent is to make decisions about whether to allocate resources to implement the intervention again or on a broader scale. In all of these, the statistical analyses use between-subjects variation (between control and experimental groups) or consistency of between-subjects variations across two or more relevant variables (correlation between variables). These widely used between-subjects statistical approaches have been challenged by numerous researchers. These researchers have demonstrated differences in intra-individual and inter-individual variation that lead to different models of change in phenomenon and association between two variables at group versus individual

levels (Borkenau & Ostendorf, 1998; Borsboom et al., 2003; Cervone, 1997; Feldman, 1995; Mischel & Shoda, 1998). Researchers have pointed out problems with between-subjects variation modeling of change several decades ago (Rogosa & Willett, 1985). They have demonstrated that group-level modeling of change is an inaccurate estimation of change at the individual level and recommend theorizing individual change by modeling individual growth and then investigating systematic individual differences in growth (Rogosa, 1995).

The foregoing does not come as a surprise to practitioners. We cannot expect the same causal mechanism and effect of intervention to work in the same way for each individual in even the most rigorous experimental (or quasi-experimental) design. For example, consider a technology-based instructional method that is intended to improve student learning and give students a chance to pace their learning. There are many factors that may affect to what extent students will benefit from such an educational intervention. Some factors include familiarity with technology, level of scaffolding needed to provide support for learning, language used in the learning software, and cultural context of the intervention. Winne (2006) demonstrates significant challenges for making claims about effectiveness of educational interventions based on group level outcomes in experimental designs. At the core of these challenges lie two axioms: (a) learners construct knowledge and (b) learners are agents. Winne argues that even if interventions are effective, learner agency introduces variance that is not accounted for in the experimental design.

Similarly, group level statistical modeling of the effectiveness of interventions or change in outcomes due to intervention can potentially overlook the positive (and the opposite of) effects of the intervention for some students. Currently, one of the most widely used applications of group-level modeling of change exists in accountability models, in particular the widely used value-added models. Generalization claims with respect to effectiveness of programs or teachers based

on such models suffer from the problems Rogosa and his colleagues elaborated on almost three decades ago. Similar to proponents of inter-individual research designs, Winne (2006) suggests that there is a need to examine individual student learning traces using interactive learning software such as the *gStudy* to inform reform efforts to improve learning.

Another widely used research approach draws on correlational studies to determine factors that are associated with better educational outcomes. Very commonly used correlational research uses statistical methods that typically employ ecological correlations (Robinson, 1950), such as Pearson correlations, which capture associations between variables for groups. These correlations use marginal frequencies for estimating group level associations. An alternative statistic individual correlation is defined as "a correlation in which the statistical object or thing described is indivisible" (p. 351). Individual correlation is based on individual level variable values such as gender, height, education level, rather than marginal frequencies for groups. Robinson demonstrates that ecological correlation differs by level of aggregation and that ecological correlations cannot be used as indicators of individual correlations. Some researchers argue that accounting for within-group heterogeneity by multi-level modeling in correlational research may address the problems of ecological correlation and individual correlations may not be needed (Subramanian, Jones, Kaddour, & Krieger, 2009). This rationale against individual correlations is not convincing to some researchers (Oakes, 2009). First, multi-level models have several assumptions that are often not met by real data. Second, multi-level models are targeted to address group-level associations and do not capture associations for individuals or sub-groups, which may have very different associations (Oakes, 2009).

The issue of heterogeneity poses itself differently in essentialist generalization. This is so because this form of generalization inherently acknowledges and is based

on the diversity in which a generalization manifests itself (see Figure 3). Read from left to right, the figure exemplifies how a generalization leads to the diversity of particulars inherent in it but not to the particulars of other generalizations (Vygotsky, 1971). The problem lies in the identification of the generalization to which the particular case of interest belongs. Thus, for example, Piaget's work on reasoning is problematic not because he did not generalize; rather, it is problematic because it does not apply in the case of the fundamental restructuring of reasoning that (schooling) culture and language bring about (e.g., Harris, 2001; Luria, 1976). Once a true generalization has been found, however, it will apply to every case; it only manifests itself differently in different cases. Therefore, in contrast to the two other forms of generalization, essentialist generalization inherently addresses heterogeneity as long as we take into account the contextual particulars relevant to the manifestation of the generalization.

Uses of Knowledge Claims as Criterion for Generalizing

A study on research use suggests that there tends to be a lack of uptake of research evidence on the part of teachers (Williams & Coles, 2003). The study shows that links between research output and practice often are not apparent. Moreover, often overlooked in the research on knowledge use is the relation between knowledge and interests (e.g., Habermas, 2008). Thus, as the introductory quotation from Austin shows, the truth or falsehood of statements (knowledge claims) depends on the intents and purposes (i.e., uses) of a statement (knowledge claim). Similarly, the question about the extent to which we can generalize research results cannot be limited to evaluating consistency, reliability across observations, or validity of interpretations (Bachman, 2009). Rather, the evaluation of the extent to which research claims are generalizable needs "to consider the uses that may be

made of our research results, and the consequences of these uses for various individuals who may be affected by them" (p. 127). Granting councils around the world already are sensitive to the relationship between knowledge and use. Thus, for example, the Canadian Institute for Health Research

> defines a knowledge-user as an individual who is likely to be able to use the knowledge generated through research to make informed decisions about health policies, programs and/or practices. A knowledge-user's level of engagement in the research process may vary in intensity and complexity depending on the nature of the research and his/her information needs. A knowledge-user can be, but is not limited to, a practitioner, policy-maker, educator, decision-maker, health care administrator, community leader, or an individual in a health charity, patient group, private sector organization, or media outlet. (CIHR, 2011)

There now exists extensive empirical evidence that knowledge is situated and specific to the circumstances so that what is useful in one setting is not useful in another (e.g., Lave, 1988; Lobato, 2006; Packer, 2001; Saxe, 1991; Tuomi-Gröhn & Engeström, 2003). It may therefore not come as a surprise that some scholars refer to knowledge in the plural form, as in "situated knowledges" (e.g., Haraway, 1991). In this section, we discuss a research use argument in the light of the preceding discussion of the three forms of generalization.

The alternate levels of generalization allow us to understand that there are different ways in which change in education may be brought about. For example, much of current educational policy practice is to target tendencies such as the overall positive correlations between educational practice and learning outcomes or an increase of group level learning outcomes. This, as in analytic generalization,

does not guarantee that every individual benefit or that some treatment is in the interest of any particular individual. In fact these generalizations may overlook the potential for negative implications of educational practice or policy on sub-groups or individuals. In essentialist generalization, results are pertinent to every single case because the claim is based on what is common to *all* cases – as long as the contextual particulars are taken into account. This kind of generalization, therefore, can be thought of as having limitations because it works itself out differently when the conditions differ – just as the children from the same parents growing up in the same family have very different biological and psychological characteristics.

Which level of generalization to be reported cannot be resolved based on research method alone but also requires consideration of the future use. There is an essential relation between (knowledge) production and the use of its product (e.g., Leont'ev, 1978). Thus, "the work – what is in the process of being produced – always already lets us encounter the what-for of *its* usability" (Heidegger, 1927/1977, p. 70, our translation, original emphasis). As any other product, the processes and contexts of their production characterize knowledge claims. But knowledge claims also contain references to the future use for which – to paraphrase Heidegger – these claims are tailor-made and which therefore "'is' present as the work emerges" (p. 71). However, those readers who have spent a lot of time in primary and secondary classrooms will have heard comments about how little (quantitative) research is useful to the real, experienced needs of teachers and their students. A considerable amount of "qualitative" research is equally unhelpful, because constructs do not pertain to anything outside of the original context of the knowledge production because the predominant aim is to produce knowledge for those involved as participants. This is so, for example, with the "authenticity criteria," which were proposed to govern fourth generation evaluation (Guba & Lincoln, 1989). These criteria for the quality of research are to ensure that

stakeholders themselves understand, are stimulated to action, and empowered to change their situation. Those readers who have spent a lot of time with policy makers may have heard complaints about research that is not attempting to provide results that generalize to broader contexts. This problem of application is not limited to statistical or case-based forms of research but is pertinent to research in general. If users have good reasons to complain, this is likely so because they have been reading reports or articles intended for a different audience (users) than themselves.

The question of what is supporting evidence for a generalization therefore also is a function of the future use of the knowledge claims made. This is so because in education there are different stakeholders with differing interests and responsibilities, and they require different forms of information (knowledge) to do their job – whether they work in the Canadian North or in a state at the Gulf of Mexico. The usefulness of claims depends on the target uses of the knowledge. Thus, a minister of education may focus on the allocation of funds to areas that have been identified as specific needs. The knowledge required is supra-individual in nature, and the relevant distinctions may be those between urban, suburban, and rural school *districts*. Knowledge of the relationship between mean parental income and achievement may lead politicians to decide about making available more funding for school-based resources to those districts serving poor neighborhoods than to those in affluent neighborhoods. On the other extreme, a teacher has to know what to do with *this* student for each of the 28 or so students they have in their classrooms. Whereas the minister of education and high-level bureaucrats need to have knowledge of the kind expressed in Figure 2a, the teacher needs to know – or learn in the course of interacting with the individual student – precisely why the student is not doing as well as he is expected to according to the general relation between IQ and achievement. That is, the teacher needs to know how to deal with deviations in

Figure 2a that statisticians treat as error variance. A superintendent of schools might decide, based on the results of (quasi-) experiments to foster teaching science using a hands-on approach over lecture style approaches. She may make available funding to assist teachers in learning how to teach with this new method. All of these decisions need to be guided by different types of evidence that the resulting actions at the student, classroom, teacher and school levels will lead to improvement.

The idea that a generalization meets the needs of particular cases underlies the concept of phronesis sometimes discussed by teacher educators (e.g., Eisner, 2002) whereby the practitioner invents conduct such that the rule/law derived from generalization is violated to the minimum while satisfying the exceptional circumstances required by solicitude (Ricœur, 1990). To provide another example, general interests are distinguished from particular interests, most often represented in and by "interest groups" and the lobbyists that represent them. Effective generalization means that the interests of all interest groups are met. Is this possible? In the context of education, the *cogenerative dialogue* is one form of praxis that brings together every different stakeholder group – e.g., students, teacher, department head, and assistant principal – for the purpose of making decisions about concrete next steps that are in the interest of *all* those using and being affected by the decisions (Roth & Tobin, 2002; Tobin, 2009).

Consideration of a set of generalizations at different levels, individual, sub-population and population, therefore occur at the very heart of educational praxis, whereby all stakeholders commit to act in the *general* interest rather than in the *particular* interests of one or the other special (interest) group. Knowledge underlying the common plan inherently is shared and therefore of generalized nature rather than of a nature particular to an individual or group. Responding to our rhetorical question, yes, it is possible to produce useful generalizations *if these*

are tailored beforehand to the needs of the particular user. Educational researchers therefore need to include the uses in their evaluations of research generalization in addition to evaluating consistency, reliability, or validity. Our recommendation thereby is consistent with the suggestion that research should be concerned with tactical authenticity by providing stakeholders with the means that allow them to empower themselves (Guba & Lincoln, 1989); but we extend this argument beyond the particular epistemological underpinnings to which it was initially applied and to all forms of generalization discussed in this article.

Final Note

The purpose of this paper is to provide an overarching framework that includes population heterogeneity and uses of knowledge as integral aspects in the process of research generalization and in the production of evidence on which educational policy analysis, evaluation, and decision-making are based. The power of research derives from the fact that it produces knowledge that can be used in multiple settings. In educational research, however, the question too often has been more about the use of qualitative or quantitative method rather than about the potential of research to contribute to the improvement of education. Yet, to paraphrase Bourdieu for our own purposes, educational research "is something much too serious and too difficult to allow ourselves to mistake scientific *rigidity*, which is the nemesis of intelligence and invention, for scientific *rigor*" (Bourdieu, 1992, p. 227, original emphasis). Mistaking rigidity and rigor would dismiss some research methods and lead us to miss out on the "full panoply of intellectual traditions of our discipline and of the sister disciplines of anthropology, economics, history, etc." (p. 227).

The problems deriving from over-generalizing exist in both quantitative and qualitative research. It is such over-generalizing that we need to guard against most vigorously by taking into account (a) the diversity in the populations of interest and (b) uses of knowledge from educational research as indicators of the quality of empirical evidence for policy and practice. Here we argue for the inclusion of population heterogeneity and knowledge uses when considering educational research generalization. With respect to the latter, one may only speculate about the absence of uses as a criterion. It may well be that the research communities represented in journals and authors of journal articles hope to reach the widest audience possible and therefore generalize their findings across specific uses. However, the different knowledge interests and needs that characterize teachers, politicians, evaluators, analysts, policymakers, or high-level administrators should highlight the importance that knowledge use is an important dimension of its generality. Including population heterogeneity as a criterion of the extent to which it is possible to generalize research findings simply means recognizing (a) diversity along a virtually infinite number of dimensions within society and (b) that what is beneficial for one identifiable group may be neutral or detrimental for another group even though they appear to be very similar. This recognition needs to be accompanied with clarity in how research findings are reported and an explicit identification of limits of generalizations. These can include clear identification of specifics of the domain about which the research question is asked including units (U), treatments (T), observing operations (O), and settings (S) of UTOS (Cronbach, 1982). We refer to these as *referents* in research reporting (Roe, 2012). In addition to descriptions of UTOS, there is a need to consider and discuss the degree to which research findings would be invariant in contexts not represented by UTOS. These will constitute the boundary conditions for the research claims.

What are the uses of knowledge claims? Are these always obvious to researchers? To understand the relation between knowledge claims and use, future research on this issue might draw on reader response theory (e.g., Fish, 1980) and other theories that emphasize the author/reader interactions (e.g., Derrida, 1988). Any statement is true and appropriate for some uses (intends, purposes) while it is false for others (Austin, 1962/1975). Thus, researchers must not only consider content and form of their communication but also the use (intent, purpose) that is to be made with statements (knowledge claims). However, they need to be explicit about the targeted levels of knowledge claims that are appropriate.

The proposed criticisms of research generalizations have implications on how research is conducted and research findings are summarized. The main limitation in analytic generalization is that it does not provide evidence of a causal link for sub-groups or individuals. In addition to making explicit the uses that the knowledge claims may be targeting, there is a need for some changes in how research is conducted. This includes

1. A need for research to demonstrate mechanisms of causality. This may help move the field in the direction of understanding for whom and how the intervention may work.
2. A need to describe intervention outcomes in three ways: positive, negative and neutral outcomes. These descriptions need to be accompanied with which individuals fall under these different categories.
3. Latent class analysis accompanied with discriminant analysis profiling latent subgroups that constitute heterogeneity.

The main criticism of probabilistic generalization is that it may not apply to sub-groups and may have limited value for guiding policy and practice. This highlights a need for defining grouping variables by intended uses of knowledge claims. To elaborate consider international large-scale assessments of reading literacy which

demonstrated average lower achievement for boys. If the intended purpose for knowledge claims, for example, is to improve low performance for boys, then the overall group results may not provide much guidance to inform practice and policy which has such a purpose. A focus of knowledge claims on the targeted population would require identifying low performing boys, examining their profiles, and determining why they are performing lower. Factors to explore could include, opportunity to learn, language of the test or the language competency of the student.

With respect to essentialist generalization, there currently exist too few studies of the qualitative kind attempting to identify invariants that hold across the range of relevant situations. The problem arises from the fact that studies identify the various manifestations of a phenomenon (e.g., different kinds of queues), which differ across people, settings, and contexts. Yet if researchers were to study the underlying work that produces the manifestations, then not only would the phenomenon itself (e.g., the work required for producing a queue) be understood but also the various manifestations (Garfinkel, 1967). As we note above, queues exist in many ways but despite the variation in the manifestation of queuing, the underlying work of queuing is actually the same (Garfinkel, 2002). With respect to education, there already exist examples of research in which the study of individual cases, generally identifying the work of producing social structure, have led to the identification of patterns that can be found in many situations within and across countries. Thus, for example, the acronym *IRE* – *i*nitiation, *r*esponse, *e*valuation – refers to a turn-taking ritual in which teachers take the first and third position (initiating, evaluating) and students take the middle position (e.g., Lemke, 1990). Its function is to mark in classroom processes culturally accepted and rejected forms of knowledge (e.g., Poole, 1994; van Eijck & Roth, 2010), thereby allowing the reproduction and objectivity of the sciences and mathematics (Roth & Gardner, 2012). Other case studies provide evidence of how, because every interview uses

language, student (mis-) conceptions inherently are cultural (common, general) rather than personal (singular, special) phenomena (Roth, Lee, & Hwang, 2008). This has far-reaching implications in that this research suggests the impossibility of "eradicating misconceptions," a long-held ideal of many science educators working from a conceptual change perspective.

Notes

[1] Some more radical "constructivist" educators favor the term "transportability" of findings (e.g., Guba & Lincoln, 1989); but the underlying concern is the same: making use of research findings in a setting other and therefore wider than where they are originally produced.

[2] The adjective "essentialist" is based on Vygotsky's (1927/1997) description of this form of generalization, which, as shown below, has as its goal "not a systematic exposition of a psychological theory . . . but precisely *the analysis of the processes in their essence*" (p. 319, original emphasis, underline added).

[3] The *design experiment* is a research method that combines experimental and case-based methods to investigate complex interventions; it is intended to produce generalizations while being useful to the particular case (Brown, 1992).

[4] In fact, there exists an insistence on the part of many "qualitative" researchers that their research ought not pursue generalization because "[t]he trouble with generalizations is that they don't apply to particulars" (e.g., Lincoln & Guba, 1985, p. 110).

[5] UTOS refers to domain about which the research question is asked, involving units (U), treatments (T), observing operations (O), and settings (S). UTOS* refers to the specific situation or class about which a conclusion is wanted (Cronbach, 1982).

[6] In the context of classical logic, this statement may sound contradictory. In dialectical logic, however, a thing is not self-identical so that it will never manifest itself in identical form, a wisdom also capture d in the Heraclitean observation that we can never step into the same river twice.

References

Austin, J. L. (1975). *How to do things with words* (2nd ed.). Cambridge, MA: Harvard University Press. (First published in 1962)

Bachman, L. F. (2009). Generalizability and research use arguments. In K. Ercikan & W.-M. Roth (Eds.), *Generalizing from educational research: Beyond qualitative and quantitative polarization* (pp. 127–148). New York, NY: Routledge.

Borkenau, P., & Ostendorf, F. (1998). The big five as states: How useful is the five factor model to describe intraindividual variations over time? *Journal of Research in Personality, 32,* 202–221.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219.

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, IL: University of Chicago Press.

Brewer, D. J., Fuller, B., & Loeb, S. (2010). Editor's introduction. *Educational Evaluation and Policy Analysis, 32*, 3–4.

Brown, A. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2*, 141–178.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. *New Directions for Program Evaluation, 31,* 67–77.

Canadian Institute for Health Research (CIHR). (2011). More about knowledge translation at CIHR. Accessed February 23, 2012 at http://www.cihr-irsc.gc.ca/e/39033.html

Cervone, D. (1997). Social–cognitive mechanisms and personality coherence: Self-knowledge, situational beliefs, and cross-situational coherence in perceived self-efficacy. *Psychological Science, 8,* 43–50.

Clark, L., & Preece, R. (2012, August 16). Boys are top of the class! Teenagers celebrate as they get A-level marks . . . and lads do better than girls at getting A* grades. *Mail Online*. Accessed December 18, 2012 at http://www.dailymail.co.uk/news/article-2188974/A-Level-Results-Day-2012-Boys-better-girls-achieving-A-grades.html

Cohen-Vogel, L. (2011). "Staffing to the test": Are today's school personnel practices evidence based? *Educational Evaluation and Policy Analysis, 33,* 483–505.

Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., Talbert, J. E. for the Stanford Aptitude Seminar (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.

Derrida, J. (1988). *Limited inc*. Chicago, IL: University of Chicago Press.

Eisenhart, M. (2009). Generalization from qualitative inquiry. In K. Ercikan & W.-M. Roth (Eds.), *Generalizing from educational research: Beyond qualitative and quantitative polarization* (pp. 51–66). New York, NY: Routledge.

Eisner, E. W. (2002). From episteme to phronesis to artistry in the study and improvement of teaching. *Teaching and Teacher Education, 18,* 375–385.

Ercikan, K. (2009). Limitations in sample to population generalizing. In K. Ercikan & M-W. Roth (Eds.), *Generalizing in educational research: Beyond qualitative and quantitative polarization* (pp. 211–235). New York, NY: Routledge.

Ercikan, K., & W.-M. Roth (2009). *Generalizing from educational research: Beyond qualitative and quantitative polarization.* New York, NY: Routledge.

Ercikan, K., Roth, W-M., Asil, M. (in press). Cautions about uses of international assessments. *Teachers College Record*.

Ercikan, K., Roth, M., Simon, M., Lyons-Thomas, J., & Sandilands, D. (in press). Assessment of linguistic minority students. *Applied Measurement in Education*.

Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology, 69,* 153–166.

Finfgeld-Connett, D. (2010). Generalizability and transferability of meta-synthesis research findings. *Journal of Advanced Nursing, 66,* 246–254.

Firestone, W. A. (1993). Alternative arguments for generalizing from data as applied to qualitative research. *Educational Researcher*, *22* (4), 16–23.

Fish. S. (1980). *Is there a text in this class? The authority of interpretive communities*. Cambridge, MA: Harvard University Press.

Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Garfinkel, H. (2002). *Ethnomethodology's program: Working out Durkheim's aphorism*. Lanham, NY: Rowman & Littlefield.

Guba, E., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Habermas, J. (2008). *Erkenntnis und Interesse* [Knowledge and interests]. Hamburg, Germany: Felix Meiner.

Haraway, D. J. (1991). *Simians, cyborgs and women: The reinvention of nature*. New York, NY: Routledge.

Harris, P. L. (2001). Thinking about the unknown. *TRENDS in Cognitive Sciences, 5,* 494–498.

Heidegger, M. (1977). *Sein und Zeit* [Being and time]. Tübingen, Germany: Max Niemeyer. (First published in 1927)

Holzkamp, K. (1983). *Grundlegung der Psychologie* [Laying the foundation of psychology]. Frankfurt/M, Germany: Campus.

Il'enkov, E. (1982). *Dialectics of the abstract and the concrete in Marx's Capital*. Moscow, Russia: Progress.

Kennedy, M. M. (2008). Contributions of qualitative research to research on teacher qualifications. *Educational Evaluation and Policy Analysis, 30,* 344–367.

Lave, J. (1988). *Cognition in practice: Mind, mathematics and culture in everyday life*. Cambridge, UK: Cambridge University Press.

Lemke, J. L. (1990). *Talking science: Language, learning and values*. Norwood, NJ: Ablex.

Leont'ev, A. N. (1978). *Activity, consciousness, personality*. Englewood Cliffs, NJ: Prentice Hall.

Leontyev, A. N. (1981). *Problems of the development of the mind*. Moscow, Russia: Progress.

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

Lobato, J. (2006). Alternative perspectives on the transfer of learning: History, issues, and challenges for future research. *Journal of the Learning Sciences 15,* 431–449.

Luke, A. (2009). Critical realism, policy, and educational research. In K. Ercikan & W.-M. Roth (Eds.), *Generalizing from educational research: Beyond qualitative and quantitative polarization* (pp. 173–200). New York, NY: Routledge.

Luria, A. (1976). *Cognitive development: Its cultural and social foundations*. Cambridge, MA: Harvard University Press.

Mannheim, K. (2004). Beiträge zur Theorie der Weltanschauungs-Interpretation [Contributions to the theory of worldview interpretation]. In J. Strübing & B. Schnettler (Eds.), *Methodologie interpretativer Sozialforschung: Klassische Grundlagentexte* (pp. 103–153). Konstanz, Germany: UVK.

Marx, K. [Engels, F.] (1962). *Werke 23: Das Kapital* [Works 23: Capital]. Berlin, Germany: Dietz. (First published in 1867)

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33* (2), 3–11.

Mischel, W., & Shoda, Y. (1998). Reconciling processing dynamics and personality dispositions. *Annual Review of Psychology, 49,* 229–258.

Molenaar, P. C. M. (1999). Longitudinal analysis. In H. J. Adèr & G. J. Mellenbergh (Eds.), *Research methodology in the life, behavioural and social sciences* (pp. 143–167). Thousand Oaks, CA: Sage.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*, 201–18.

Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2003). The relationship between the structure of inter-individual and intra-individual variability: A theoretical and empirical vindication of developmental systems theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development* (pp. 339–360). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Moss, P. A. (2012). Exploring the macro-micro dynamic in data use practice. *American Journal of Education, 118*, 223–232.

Oakes, W. S. (2009). Individual, ecological and multilevel .fallacies. *International Journal of Epidemiology, 38*, 361–368.

Oliveri, M. E., Ercikan, K. & Zumbo, B.D. (in pressa). Accuracy of DIF detection methods for heterogeneous groups. *Applied Measurement in Education.*

Oliveri, M. E., Ercikan, K., & Zumbo, B. (in pressb**).** Analysis of sources of latent class DIF in international assessments*. International Journal of Testing*.

Packer, M. (2001). The problem of transfer, and the sociocultural critique of schooling. *Journal of the Learning Sciences, 10*, 493–514.

Poole, D. (1994). Routine testing practices and the linguistic construction of knowledge. *Cognition and Instruction, 12*, 125–150.

Ricœur, P. (1990). *Soi-même comme un autre* [Oneself as another]. Paris, France: Seuil.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357.

Roe, R. (2012, July). Using referents to improve generalization in psychological research. Paper presented at the symposium 'Limits of Generalizing in Psychological Research' at

the 30th International Congress of Psychology Cape Town, South Africa.

Roderick, M. (2012). Drowning in data but thirsty for analysis. *Teachers College Record, 114* (11).

Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal research," plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–65). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika, 50*, 203–228.

Roth, W.-M. (2009a). Phenomenological and dialectical perspectives on the relation between the general and the particular. In K. Ercikan & W.-M. Roth (Eds.), *Generalization in educational research* (pp. 235–260). New York, NY: Routledge.

Roth, W.-M. (2009b). Specifying the ethnomethodological "what more?" *Cultural Studies of Science Education, 4*, 1–12.

Roth, W.-M. (2012). *First person methods: Toward an empirical phenomenology of experience*. Rotterdam, The Netherlands: Sense Publishers.

Roth, W.-M., & Gardner, R. (2012). "They're gonna explain to us what makes a cube a cube?" Geometrical properties as contingent achievement of sequentially ordered child-centered mathematics lessons. *Mathematics Education Research Journal, 24*, 323–346.

Roth, W.-M., Lee, Y. J., & Hwang, S.-W. (2008). Culturing conceptions: From first principles. *Cultural Studies of Science Education, 3*, 231–261.

Roth, W.-M., McGinn, M. K., Woszczyna, C., & Boutonné, S. (1999). Differential participation during science conversations: The interaction of focal artifacts, social configuration, and physical arrangements. *Journal of the Learning Sciences, 8*, 293–347.

Roth, W.-M., & Tobin, K. (2002). *At the elbow of another: Learning to teach by coteaching*. New York, NY: Peter Lang.

Saxe, G. B. (1991). *Culture and cognitive development: Studies in mathematical understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Slavin, R. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluation. *Educational Researcher, 37* (1), 5–14.

Snow, R. E. (1989). *Aptitude-treatment interaction as a framework for research on individual differences in learning*. New York, NY: W. H. Freeman.

Song, M., & Herman, R. (2010). Critical issues and common pitfalls in designing and conducting impact studies in education: Lessons learned from the What Works clearinghouse (phase I). *Educational Evaluation and Policy Analysis, 32*, 351–371.

Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Saenz, L., Yen, L., Fuchs, L. S., & Compton, D. L. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis, 30,* 368–388.

Subramanian, S. V., Jones, K., Kaddour, A., & Krieger, N. (2009). Revisiting Robinson: The perils of individualistic and ecologic fallacy. *International Journal of Epidemiology, 38*, 342–360.

Tobin, K. (2009). Repetition, difference, and rising up with research in education. In K. Ercikan & W.-M. Roth (Eds.), *Generalizing from educational research: Beyond qualitative and quantitative polarization* (pp. 149–172). New York, NY: Routledge.

Tuomi-Gröhn, T., & Engeström, Y. (2003). Conceptualizing transfer: From standard notions to developmental perspectives. In T. Tuomi-Gröhn & Y. Engeström (Eds.), *Between school and work: New perspectives on transfer and boundary-crossing* (pp. 19–38). New York, NY: Pergamon.

Turner, E. O., & Coburn, C. E. (2012). Interventions to promote data use: An introduction. *Teachers College Record, 114* (11).

U. S. Department of Education. (2004). *Two years of accomplishment with No Child Left Behind*. Accessed February 27, 2011 at http://www2.ed.gov/news/pressreleases/2004/01/01082004factsheet.html

van Eijck, M., & Roth, W.-M. (2011). Cultural diversity in science education through novelization: Against the epicization of science and cultural centralization. *Journal of Research in Science Teaching, 48*, 824–847.

Vygotsky, L. S. (1971). *Psychology of art* (Scripta Technica, Transl.). Cambridge, MA: MIT Press.

Vygotsky, L. S. (1927/1997): The historical meaning of the crisis in psychology: A methodological investigation. In W. R. Rieber & J. Wollock (Eds.), *The collected work of L. S. Vygotsky vol. 6* (pp. 233–343). New York, NY: Kluwer Academic / Plenum Publishers.

Williams, D., & Coles, L. (2003). The use of research by teachers: Information literacy, access and attitudes. Research Report 14. Department of Information Management, The Aberdeen Business School. Accessed January 22, 2013 at www.rgu.ac.uk/files/ACF2B02.pdf

Winne, P. (2006). How software technologies can improve research on learning and bolster school reform . *Educational Psychologist, 41*, 5–17.

Winne, P. H., & Nesbit, J. C. (2010). The psychology of academic achievement. *Annual Review of Psychology, 61*, 653–678. doi:10.1146/annurev.psych. 093008.100348

Yin, R. K. (2008). *Case study research: Design and methods* (vol. 5). Thousand Oaks, CA: Sage.

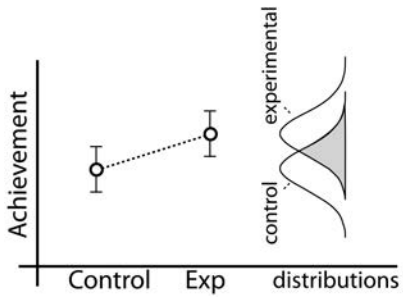

Figure 1. Comparisons between achievement of a control and an experimental group following a hypothetical experiment.

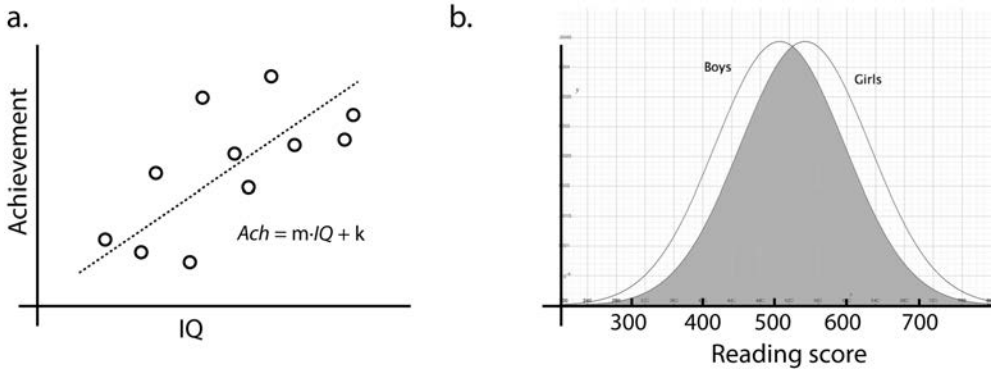

Figure 2. a. In correlational research, a trend is observed and generalized to the population. b. PISA 2009 reading results for Canadian boys and girls. The grey represents presence of boys who score higher than some of the girls and girls who score lower than some boys.

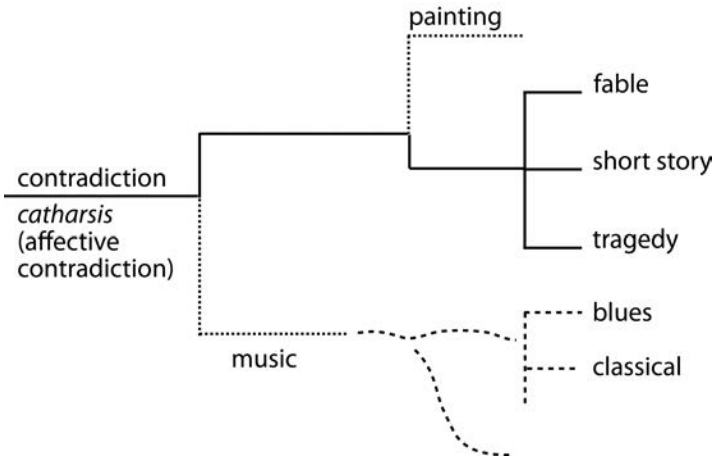

Figure 3. Pattern of Vygotsky's (1971) derivation of the psychology of art (contradiction of emotions that move in two opposing directions). The generalization is true *for every case*.