In *The Best Available Evidence: Decision Making for Educational Improvement*, the editors and contributing authors explore the intricacies of working with data and evidence for the purpose of organizational development in educational institutions. A broad theme that runs throughout this book is the need for policy makers and practitioners to be informed and critical consumers of educational research. The chapters in this volume explore quantitative, qualitative, narrative, and practitioner research approaches and explore the implications for evidence use in educational improvement efforts.

Many current texts provide an instrumental resource for educational leaders for use in designing road maps for improvement. As such, these texts offer a perspective based on assumptions that educational personnel are the recipients of predetermined knowledge and evidence, and it is the task of instructors and teachers to implement received knowledge of “best practice”. In this book, we suggest that teachers, instructors, educational leaders, and policy makers are equally engaged in the creation of knowledge and the establishment of improvement objectives. Further, we address questions concerning what constitutes improvement, how practitioners and policy makers can assess the utility and veracity of evidence, and how evidence might be considered in productive and ethical ways. This volume is intended for a broad readership of teachers, post-secondary instructors, graduate students, educational leaders, and policy makers. Finally, this book will combine K-12 perspectives on educational improvement with perspectives from the research on post-secondary improvement.
The Best Available Evidence
The Best Available Evidence

Decision Making for Educational Improvement

Edited by

Paul Newton and David Burgess

University of Saskatchewan, Canada
## TABLE OF CONTENTS

| Preface | vii |
| Acknowledgements | xi |
| 1. Evidence-Driven Policy and Practice in Education: Issues of Veracity in Quantitatively Based Research | 1 |
  *José Da Costa* |
| 2. Issues with Evidence in Constructivist/Interpretivist Educational Research | 19 |
  *Bonnie Stelmach* |
| 3. Exploring Types of Educational Action Research: Implications for Research Validity | 33 |
  *Paul Newton and David Burgess* |
| 4. Narrative, Insight, and Effective Decision Making in Education | 47 |
  *Derek Stovin* |
| 5. Values, Virtues, and Human Capabilities: A Positive Perspective on Educational Improvement | 63 |
  *Sabre Cherkowski and Keith Walker* |
| 6. The Cartography of Success: Mapping the Education Data Terrain toward Ethical Practice for Continuous Improvement | 81 |
  *Scott Tunison* |
| 7. Learning Organisations and Their Relationship to Educational Improvement | 99 |
  *Pamela Timanson and José Da Costa* |
| 8. The Assessment Imperative: Evidence-Based Decision Making in Student Affairs | 113 |
  *Vicki Squires* |
| 9. Change and Improvement in Post-Secondary Education | 129 |
  *Robin Mueller* |
| 10. Decision Making and Problems of Evidence for Emerging Educational Technologies | 147 |
  *Erika E. Smith and Richard Hayman* |
<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Epilogue: Evidence and the Research-Practice Gap in Education</td>
<td>167</td>
</tr>
<tr>
<td><em>Paul Newton and David Burgess</em></td>
<td></td>
</tr>
<tr>
<td>About the Contributors</td>
<td>175</td>
</tr>
<tr>
<td>Index</td>
<td>179</td>
</tr>
</tbody>
</table>
This volume is born from a landscape where educational improvement efforts are being directed by calls for evidence—evidence that is commonly presented as non-problematic and representative of a unitary corpus of “what we know.” Throughout this book, we characterize three broadly conceived classes of knowledge and evidentiary claims within contemporary educational improvement discourse: (a) explanatory—primarily quantitative approaches; (b) interpretive/constructivist—primarily qualitative approaches; and (c) practitioner-driven—primarily participatory approaches (Guba & Lincoln, 2005). Practitioners and policy-makers are confronted with multiple forms of educational research to consider. We hope that this edited volume will add to the conversation with respect to the use of research and evidence in educational improvement in pre K-12 and post secondary contexts. The authors in this volume explore the potential, intricacies, and limitations of a number of research approaches and the implications for the application of research evidence in education.

In Chapter 1, José da Costa begins by establishing that evidence from quantitative research requires a sophisticated understanding on the part of consumers of research about concepts of significance, importance, and the validity of research findings. He argues that educational policy makers and practitioners require a significant ability to understand “threats to validity” in quantitative research, and he suggests potential methods to ensure more mature and informed approaches to research-based policy making in education.

In Chapter 2, Bonnie Stelmach examines the challenge of transferability or generalizability of qualitative academic research to address the needs of large-scale reform and improvement agendas. She begins her chapter with a discussion of the confusion between constructionist and interpretivist qualitative research paradigms. While acknowledging the limitations of qualitative methodologies, she rejects attempts to measure the veracity of knowledge claims by holding such methods to the standards of quantitative approaches—paying particular attention to the evidence-based education movement and such initiatives at the What Works Clearinghouse.

Chapter 3 includes a discussion of practitioner research and action research with respect to issues of validity. Paul Newton and David Burgess reconceptualize validity as a contingent concept dependent on the purposes of the research. In this chapter, they present three action research modes that call for different approaches to validity and the justification of knowledge claims.

Derek Stovin, in Chapter 4, discusses the potential for narrative forms of research to inform educational improvement. He begins by providing a description of narrative approaches within qualitative research followed by the presentation of
examples of the use of narrative approaches in educational research. Finally, Stovin explores the potential for narrative research to inform decision making and policy making in education.

In Chapter 5, Cherkowski and Walker propose alternate ways of measuring schools for the purpose of improvement. They draw on literature from “positive psychology” to assert that happiness and well-being are central aims for educational improvement efforts. They suggest that traditional school improvement efforts are based in deficit models, and they explore the potential for appreciative approaches. They conclude with a framework for understanding schools as sites of human flourishing and explore the implications for educational leaders.

In Chapter 6, Scott Tunison outlines some of the potential risks inherent in data-driven accountability systems and suggests a robust set of ethical practices for data use and deployment. He argues that foundational to ethical data practices are the notions of relationships, mutual trust, and respect.

Chapter 7 is a venue for Pamela Timanson and José da Costa’s look at the literature concerning organizational learning, learning organizations, and learning communities in organizational studies and educational administration. They identify the potential for restructuring school as learning organizations and explore the tensions and contradictions of using such structural lenses to improve schools. In particular, they challenge the notion of using the learning organization concept to leverage system-level goals and priorities advocated in the interests of educational improvement and reform, and second that educational reform and localized educational improvement efforts are historically linked with current emphases on accountability, new public management, and standardized testing in the English-speaking world.

In Chapter 8, Vicki Squires explores the role of data, accountability, and institutional assessment in post-secondary institutions. She discusses the emerging practices of assessment (particularly in the student services areas) and its role in institutional planning and decision making. She uses several examples of assessment projects to illustrate the principles of effective data use in post-secondary institutional improvement efforts.

In a similar vein, Robin Mueller explores change and improvement in post-secondary education in Chapter 9. In this chapter, she reviews the history of universities and examines current trends in the evolution of post-secondary institutions. Finally, Muller suggests frameworks for thinking about change efforts in the post-secondary sector.

In Chapter 10, Erika Smith and Richard Hayman review issues facing post-secondary institutions with respect to emerging technologies. They explore the problem of the rapidity of change in the area of emerging technologies and the challenges this presents in acquiring timely and relevant evidence for change efforts. In this chapter, they suggest potential practices for addressing the evidence gap for decision makers in post-secondary contexts.
We would like to echo the comments of our contributors in suggesting that practitioners, graduate students, educational leaders, and policy makers would benefit from a level of sophistication as “readers” of the best available evidence. This evidence-informed decision making is dependent upon (a) an ability to determine the relevance and utility of evidence, (b) an ability to determine the strengths and limitations of the various forms of evidence, and (c) an understanding of the ethical and moral implications of evidence and data.
ACKNOWLEDGEMENTS

We would like to offer our sincere thanks to our colleagues, family, and students for their support as we worked through this edited book. Our colleagues in the Department of Educational Administration at the University of Saskatchewan (past and present) have been central in our development as scholars and have offered encouragement throughout this and our other scholarly endeavors. To our current and former graduate students, your work provides the impetus to contemplate the ways in which research and evidence contribute to the improvement of educational institutions. Also, regularly overlooked when we seek to offer gratitude, our colleagues who work tirelessly as practitioners and leaders in educational institutions. We owe a debt of gratitude to you for your diligent work in the field, and we owe a debt of responsibility to reflect more often on the ways in which research and evidence might be produced that can make a real difference in the lives of instructors, leaders, and students. Mostly, we would like to express our most sincere gratitude to our contributors, whose work represents a novel contribution to thinking about evidence and educational decision making. We also thank you for sticking with us on the longer than expected journey. Finally, to Michel Lokhorst at Sense, thank you for your advice, guidance, and patience.

Teresa, Connor, and Sean; and Liam and Lauren, thank you for your patience, stability, support, and love throughout.
1. EVIDENCE-DRIVEN POLICY AND PRACTICE IN EDUCATION

Issues of Veracity in Quantitatively Based Research

For several decades, calls have been made by a variety of researchers and practitioners for basing decisions in educational policy and practice in evidence rather than on more traditional and common approaches grounded in personal political ideology or personal preference and preconceived notions drawing on unsystematic analysis of anecdotal data. Even unsystematic interpretation of data gathered through large-scale initiatives, which are conducted very systematically (e.g., provincial or state administered achievement tests, Program for International Student Achievement – PISA, Trends in International Mathematics and Science Study – TIMSS), often result in invalid conclusions and invalid courses of action based primarily in preconceived ideologies. Interestingly, despite the numerous calls for grounding decision making in data, policy development often still seems to be based on selective application of research supporting decision-maker predispositions.

Basic types of rationalistic research fall into three main categories: descriptive, correlational, and experimental. As the name implies, descriptive research provides descriptions, typically drawing on descriptive statistics (e.g., mean, median, mode, standard deviation, range, etc.) to give readers a sense of the context being studied. Correlational research focuses on how two or more variables change together, either in direct or inverse relationships. The most basic measures of correlation rely on Spearman (non-parametric test of association between two variables), Kendall (non-parametric test of dependence between two variables), and Pearson (parametric test of association between two variables) correlation coefficients or derivatives of these. Experimental research and, the closely related, but methodologically more problematic, quasi-experimental research, allows researchers to explore how variables are related to each other, either directly or inversely, just as with correlational research, and it allows researchers to make claims about cause/effect relationships. Educational policy makers and practitioners often read any of these types of research and begin, erroneously, imagining how to enact change that will lead to desired outcomes, whatever they might be. Even educational policy makers and practitioners who recognize the fallacies of drawing on descriptive or correlational research to make changes to their context, may make inappropriate decisions with regard to practices and policies when outcomes from carefully controlled settings are applied in field settings which do not match the experimental setting contexts. Decisions
of this type are often based on imagining a desired outcome described in research, then ignoring many contextual factors while implementing the method described in the research gathered (e.g., the current trend to try to mimic the outcomes of Finnish schools, as described by people such as Salhberg, 2011 without heeding his warnings of the political and social complexity of the system).

This non-systematic and biased approach may be exacerbated by additional issues of validity and reliability of research conducted in field settings (i.e., classroom or school settings using ex post facto and quasi-experimental designs). As anyone who has conducted field-based research knows, classrooms and schools are not ideal laboratory settings where extraneous variables can be carefully controlled, or indeed, even monitored and noted. Such field-based research is messy and requires much replication before confidence in findings can be firmly established. Furthermore, measurement of abstract social phenomena is fraught with difficulty. Measurement in the field of education, and in the social sciences generally, attempts to link highly abstract constructs (e.g., efficacy, achievement, satisfaction, etc.) with empirical indicators. Empirical indicators form the evidence seen to be a central tenet of empirical sciences which draws on data, collected through any number of ways or instruments, either to lend credence for or refute researcher hypotheses (Popper, 1972). Confounding factors, however, can contaminate evidence and often lead to seemingly contradictory results across, what may appear to be, very similar studies. Many of these contradictory results can be traced back to issues of validity of the method and the subsequent veracity of the findings.

The fundamental purpose of this chapter is to focus on issues that should be taken into account when considering research and its applicability to research consumers’ contexts by assisting the practitioner or policy maker who desires to move beyond basing decisions on pre-dispositions or “gut instinct.” This chapter is divided into five sections, Section one deals with threats to validity of research conducted using quasi-experimental and ex post facto research designs. The second section focuses on issues created when violations of assumptions underlying statistical analyses are not reported and on issues to consider when reading research utilizing parametric statistical analyses. Section three addresses the issues inherent in traditional null-hypothesis significance testing and the need to shift to estimation and effect size measures. The fourth section explores ways of establishing and understanding the importance (as opposed to the statistical significance) of statistical findings. Having dealt with the aspects necessary to assess the quality of evidence and the data underlying the evidence, the fifth section explores the policy-making and decision-making literature specifically identifying the problem of not drawing systematically on research to inform policy and practice.

Threats to Validity

The need to understand what constitutes valid data and the ability to draw appropriate inference from those data to inform decisions must be understood not
only by researchers, but also by consumers of research. Validity in research is not new and many seminal works have been written on the topic from quantitative and qualitative perspectives. Some of the more well known include: Campbell and Stanley (1963, 1966), Cook and Campbell (1979), Guba (1981), and Lincoln and Guba (1985). Foundational to notions of what constitutes credible research for grounding decisions are four overarching (i.e., applicable to empirical qualitative and quantitative forms of research) conceptions of validity originally described by Guba (1981): (a) truth value, (b) applicability, (c) consistency, and (d) neutrality. While Guba’s 1981 paper focuses on validity in qualitative research, he provides useful conceptions proposing parallels between quantitative and qualitative research that are particularly helpful for thinking about validity broadly. More specific, but still in keeping with the broader conceptualizations of validity, to research conducted from a rationalistic paradigm was the validity framework proposed earlier by Campbell and Stanley (1963, 1966), refined later by Cook and Campbell (1979). Their framework for considering validity for research conducted in education, and the social sciences generally, is based on the need to identify potential sources of invalidity in research when intact groups of participants, or those who self-selected for participation become the source or sources from which data are gathered. Much rationalistic field-based research, just as laboratory-based research, seeks to explore and understand cause/effect relationships. The four areas of potential invalidity named are, in essence, areas that must be considered in establishing cause/effect relationships. In order to do so, researchers must be able to:

1. Establish that changes to a presumed cause (i.e., the independent variable) are related to changes in what is thought to be an effect (i.e., the dependent variable); in other words, there is co-variation between the cause and effect.
2. Ascertain temporality of the cause and effect (i.e., establish that the presumed cause did occur prior to the presumed effect).
3. Rule out plausible alternative explanations, or at least establish that the presumed cause is, truly, the most reasonable explanation for the observed outcome. In fact, when alternative explanations exist for the observed outcome, one cannot have any confidence in the presumed cause/effect relationship.

In meeting these three criteria in rationalist field-based research, specifically, researchers and research consumers must be very concerned with: (a) internal validity, (b) construct validity, (c) statistical conclusion validity, (d) external validity, (e) reliability, and (f) objectivity. While brief descriptions of these forms of validity are provided below, the reader is encouraged to read the very detailed discussions of validity offered by Stanley and Campbell (1963, 1966), Cook and Campbell (1979), and the many others whose ideas have evolved from these seminal works.

**Internal validity.** Internal validity refers to researcher control over research variables in the study setting. This includes control over variables of which the researcher may not even be aware. These variables may cause outcomes that may
be incorrectly attributed to variables of interest. Lack of researcher control over variables not central to the research at hand can lead to confounding of factors believed to cause effects observed in dependent variables. The central question researchers and research consumers must ask themselves about research findings is: what besides the independent variables of interest may have led to the results being observed?

In laboratory settings, researchers most typically rely on random assignment of subjects to treatment and control groups in order to randomly and evenly distribute all characteristics participants bring (e.g., random assignment of any characteristics which may be possessed by research participants, examples may include height, weight, intellectual capacity, etc.) between the comparison groups making them equivalent prior to administration of any treatment or implementation of any program. Field-based researchers typically do not have the benefit of being able to randomly assign participants to either treatment or control groups, these researchers most often have to work with intact groups. In schools these intact groups may consist of classes, schools, jurisdictions, or other units over which the researcher has no control with respect to who is in the group. In order to establish, at least, some degree of certainty that effects observed in a dependent variable are due to the independent variable of interest, researchers are well advised to systematically consider and rule out the threats to validity as described by Cook and Campbell (1979). These include:

1. History – something, besides the treatment, which occurs between a pre-test and a post-test that causes post-test results increase or decrease.
2. Maturation – skills develop or deteriorate due to age progression.
3. Testing – familiarity with test or questionnaire format leads to improved performance.
4. Instrumentation – observers gathering data change in systematic ways.
5. Statistical regression (i.e., regression to the mean).
6. Selection bias – participants self-select on the basis of some criterion or multiple criteria of interest of which the researcher is typically unaware.
7. Mortality of study participants (particularly problematic if certain sectors of participants drop out of the study causing groups to become biased in particular ways).
8. Interaction of history, maturation, testing, or instrumentation with selection bias.
10. Diffusion or imitation of treatment (i.e., without researcher knowledge, control or comparison group members interact with the members of the experimental group and learn about and are effected by the treatment meant only for the experimental group).
11. Compensatory equalization of treatments (i.e., the control group receives treatment conditions without the knowledge of the researcher, for example
when teachers share materials with each other because they perceive them to be beneficial).

12. Compensatory rivalry (i.e., John Henry effect).
13. Resentful demoralization of respondents receiving less desirable treatments, or no treatment, with these people purposely sabotaging study results (Cook & Campbell, 1979).

**Construct validity.** Researchers have a responsibility to carefully define the research constructs central to their research focus. These constructs and the relationships found amongst them are, usually, the focus of rationalistic research. Problems in defining the constructs or researchers’ inabilities to describe the relationships between and among the constructs can allow for rival explanations of the mechanisms underlying what is observed. If one considers such constructs as intelligence, love, efficacy, self-esteem, achievement, learning, to name a few, the complexities underlying these are great. Even with carefully crafted definitions it is still possible for others to misunderstand what the “label” means. For example, many lay people would consider learning and achievement to simply be synonyms for the same construct, however, a careful examination of these quickly reveals them as different. A variety of considerations must be attended to by researchers to ensure strong construct validity. These are:

1. Adequately defining constructs such that clear operational definitions are understood by the researcher and the consumers of the research, an operational definition is one that is worded such that it can be measured.
2. Mono-operational bias refers to a very narrow or limited operationalization of what is being measured, this is best attended to by using multiple indicators of a particular construct (e.g., operationalizing student achievement could be done through standardized tests, examples of student work, and teacher opinion).
3. Avoiding mono-method bias involves collecting indicators of a construct using a variety of methodological approaches (e.g., paper-and-pencil questionnaires, interviews, observation, etc.).
4. Hypothesis guessing in which the research participants try to guess what the researcher is looking for, then they provide the answers, usually, that the researcher is predisposed to hearing or seeing.
5. Evaluation apprehension connects to people’s general desire to be positively evaluated, often responses to questions, written or oral, will be tempered such that, as an individual, the respondent’s reputation or expertise is not called into question (e.g., teachers rate their school highly because it is a reflection on them).
6. Expectancies – researchers, regardless of the research tradition they gravitate toward, observers, supervisors, research participants, etc., all have beliefs about an almost infinite number of topics and issues, these predispositions often result, if not carefully accounted for, in findings that reflect the predispositions rather
than the actual research outcomes (e.g., researchers find what they are looking for, teachers gathering data on behalf of researchers see effects they believe they should see).

7. Confounding of constructs occurs when something, other than what is thought to be the independent variable, causes the observed effects in the dependent variable (e.g., in single group grade one classroom study, is the development of reading ability a function of a particular reading program or approach to teaching, or is it a function of an excellent teacher who knows how to provide students with appropriate scaffolding, or both?).

8. Level of constructs refers to the “amount” of the independent variable needed to cause an observable effect on the dependent variable, this is partly related to the power (i.e., the ability of the research to correctly reject the statistical null hypothesis) of the research as well as to the sensitivity, including the validity and reliability, of the measuring instruments.

9. Interaction of testing and treatment can threaten construct validity when measures influence the observed outcome of the treatment (e.g., a pre-test may provide students with clues about the “correct” answers).

10. Interactions of different treatments, which may be inadvertently introduced by the researcher or teachers in classroom field settings, may have a synergistic effect that cannot be teased apart by the researcher even if she or he becomes aware of the interaction.

11. Restricted generalizability across constructs is often problematic in field settings in education as researchers use different operational definitions of the same construct (e.g., student achievement is often measured by standardized achievement tests, but what is found to be related to this operationalization of achievement is limited since teachers are responsible for many other aspects of students’ education including socialization) (Cook & Campbell, 1979).

Statistical conclusion validity.  At its most fundamental level, statistical conclusion validity deals with the question, “is this study sensitive enough to detect a difference or a relationship if one exists?” Seven threats exist to this form of validity, namely:

1. Low statistical power (e.g., the sample size and the effect size of the quality under investigation, either singly or in combination are simply too small to be measured).
2. Violated assumptions of the statistical tests (addressed in more detail in the next section of this chapter).
3. Fishing and error rate (e.g., statistical results that occur mainly by chance).
4. Reliability and validity of data collection instruments (e.g., the measuring instruments provide data that are repeatable and reflect the construct under study; as an example, IQ tests are often accused of measuring reading ability to a large extent).
5. Reliability of the treatment implementation or data collection implementation (e.g., treatments are administered to all groups as intended by the researcher, or data collection instruments are administered under the same conditions for all participants).

6. Random irrelevancies in the research setting (e.g., in field settings research conditions are often changed, knowingly or unknowingly, by teachers, parents, administrators, etc.; even a fire alarm ringing during one group’s experience will create a random irrelevancy for the group).

7. Random heterogeneity of the respondents (e.g., the refers to the possibility that a treatment or set of questions may affect a particular type of person differently from others, this could manifest in many ways, but gender, age, etc. are some possibilities) (Cook & Campbell, 1979).

External validity. The concept of generalizability is at the heart of external validity. At issue is the ability to apply what has been learned from the participants in the specific research context to other contexts. Generalizability falls into three basic categories related to persons, settings, and times:

1. Participant selection (e.g., systematic differences between the participants in the research sample or samples and the population to which the results will be generalized).

2. Setting (e.g., differences in context between where the research is conducted and the settings to which the research will be generalized).

3. History (e.g., differences in temporal context between when the research is conducted and the time periods to which the research will be generalized) (Cook & Campbell, 1979).

As the context in which the research was conducted and the context to which the researcher is interested in generalize diverge, the less certainty there can be that the research findings will, indeed, apply to the larger context.

Violations of Statistical Assumptions and Appropriateness of Analysis Method

All statistical analyses, those that compare groups and those that explore relationships between or among variables, are founded on particular assumptions regarding the characteristics of the sample or the population from which the sample is drawn. At the most fundamental level, non-parametric analysis methods are distribution-free (i.e., they do not come from populations in which the particular characteristic is normally distributed, for example “income”) and can be used for analysis of nominal, ordinal, equal interval, and ratio data (Gall, Gall, & Borg, 2007; Gay, Mills, & Airasian, 2012). The main assumptions underlying the most fundamental non-parametric statistical analyses are that the data are collected from a random sample and that the data come from individuals that are independent of each other. However,
non-parametric statistical analyses tend to be less powerful than parametric statistical analysis techniques (i.e., they are less likely to detect a true difference if one actually exists—less likelihood of a Type II error).

Parametric statistics, on the other hand, have more underlying assumptions but are appropriate for equal interval data and ratio data, in addition to randomness and independence, the most fundamental analyses must also be based on data that are normally distributed and exhibit homogeneity of variance. More advanced non-parametric and parametric analyses will have additional assumptions that must be met in order to use them appropriately. Violations of the statistical assumptions, knowingly or unknowingly, often lead to analyses that support incorrect conclusions about the substantive question at hand.

Tests for assessing the extent to which assumptions are violated are available and should be reported by researchers. Interestingly, many research reports do not provide the reader with any assurance that assumptions underlying the statistical tests of significance have been met. Educational policy makers and practitioners are cautioned to look for assurance in research reports that statistical test assumptions have been met.

Earlier it was pointed out that parametric analyses are appropriate for equal interval and ratio data. Interestingly, many educational surveys and data collection instruments collect categorical or, at best, ordinal data (e.g., Likert-type scale gathering satisfaction data, where: 1 = “strongly disagree,” 2 = “disagree,” 3 = “neutral,” 4 = “agree,” 5 = “strongly agree”). While categorical data are often reported as frequencies and analysed using non-parametric statistics, it is not unusual to observe ordinal data being treated as though they were on an equal interval or ratio scale with means and standard deviations being reported. This is highly problematic since the research consumer can have no confidence that the differences in means and standard deviations between two groups or between two observation points from the same group are meaningful. This, of course, leads to problems of validity of the claims made in the research report.

Educational policy makers and practitioners are cautioned to look for evidence that statistical analyses are appropriate to the kinds of data gathered (i.e., categorical, ordinal, equal interval, ratio) and that the report writer has provided assurance that the underlying assumptions for the statistical tests have been met. This is not to suggest that consumers of research must be able to conduct these analyses, but they should look for the assurances before simply accepting conclusions and recommendations made by the researcher. This is also not to suggest that research reporting that an assumption underlying a statistical test has been violated should be entirely dismissed, however, it should be taken cautiously.

* A Different Way of Thinking about Significance

Traditionally, students learning about null-hypothesis significance testing (NHST) using inferential statistics were taught to select the “risk,” the alpha (α) level,¹
they were willing to take in rejecting a false null-hypothesis (i.e., type I error) and to clearly set out the statistical null hypothesis and the statistical alternate hypothesis in advance of whatever analysis they were going to conduct (e.g., Gall et al., 2007; Glass & Hopkins, 1996; Hopkins, Glass, & Hopkins, 1995). Social science research conventions typically set the alpha level at the 0.05 level (i.e., the researcher is willing to make a type I error 5 times out of every 100 times that a random sample could be drawn from a population for the research). In educational research that may have higher potential for harm, researchers may adopt a more conservative alpha level of 0.01 (i.e., theoretically the researcher is willing to make a type I error only 1 time per 100 times that a random sample could be drawn from a population for the study). For exploratory research, researchers may make a case for relaxing the alpha level to 0.10 (i.e., the researcher is willing to make a type I error 1 time in 10 times that a random sample could be drawn from a population for the research). After a priori setting the appropriate alpha level for determining when to reject or not reject the null hypothesis being tested through the inferential statistical analysis, researchers could then gather their data and conduct their analyses, ultimately finding evidence for rejecting or not rejecting the null hypothesis of difference or null hypothesis of association. Unfortunately, this sets up a dichotomy that simply paints the world as black or white: significant or not significant. This view of research is grounded in the probability logic in selecting samples from a population, significance simply leads the researcher to not reject the statistical null hypothesis or it leads the researcher to reject the statistical null hypothesis. It is simply a binary (Neyman & Pearson, 1933). While this may be a very reasonable approach in areas of research in which true random samples can be drawn from infinitely large populations, it is fundamentally flawed in research drawing on participants who volunteer and who are available to actually participate in research.

It is fairly common knowledge that in research publication, results that are non-significant (i.e., the null hypothesis is not rejected) typically do not get published. Cumming (2013) summarizes: “a ‘statistically significant’ effect in the results section becomes ‘significant’ in the discussion or abstract, and ‘significant’ shouts ‘important’” (p. 3). This in combination with other factors, for example, statistical analysis software packages that are much more user friendly and accessible to researchers whose understanding of statistical analysis theory and methodology may be somewhat lacking, has led many researchers to engage in what Cumming (2013), and others, refer to as “cherry picking”—conducting analyses in which data from some participants are left out, some variables are left out, some variables that had not been considered originally are included, in the interest of capitalizing on what are actually random differences, or chance, in datasets. These researchers really have no idea whether or not they happened to have randomly drawn (or the sample may not be random at all as highlighted by Cook and Campbell in their threats to validity) a sample that is truly representative of the population from which it was drawn, and to which they wish to generalize.
A second caution is offered here in relation to statistical tests of significance. While many statistical analysis approaches exist to ensure that researchers retain control of the alpha level when examining relationships among multiple variables or multiple groups, there are instances in the research literature in which multiple univariate analyses have been conducted, as though the data for each analysis was gathered independently from a different randomly selected sample. This leads to issues in which pre-existing relationships between multiple variables or constructs are overlooked, a factor stemming from inter-dependence of individuals from one test to the next (e.g., Hopkins et al., 1995; Glass & Hopkins, 1996; Sprinthall, 2011). Furthermore, research reports using this approach increase the chances of making a type I error (i.e., reject the null-hypothesis when it is actually true). In fact, every univariate statistical analysis adds to the probability that a type I error will be made, however, neither the researcher nor the research consumer will know in which analysis in the series of analyses this occurred.

Educational policy makers and practitioners are urged to look for clear evidence that research reports are not simply capitalizing on chance to report findings that emerged. This is true of research that reports p-values without consideration for their hypothesis-testing purpose as well as research that ignores the increased probability of finding a “false positive” result. Having identified two very important drawbacks of NHST, the question of what alternatives are possible to address these problems remains. Literature emerging over the past decade may offer a way out by focusing on transparent reporting of research, analyses that emphasize importance of findings over significance, and replication of research to enable multiple approaches for establishing the robustness of research findings

Research Transparency, Importance, Cumulative Evidence

As highlighted above, dissatisfaction with NHST approaches have been surfacing in various research communities. Fidler (2010) points, in particular, to the dissatisfaction with Null Hypothesis Significance Testing amongst psychologists. In fact the 1994 edition of the American Psychological Association (APA) Publication Manual calls for reporting statistical power along with effect size statistics of statistical analyses. This guidance as well as an alternate recommendation to provide confidence intervals related to reported statistical analyses is provided in the most recent edition of the APA Publication Manual (2010). In addition to the approximately five-dozen APA published journals, many other journals, not associated with APA, also make use of the APA Style Manual for article submissions. Psychological Science, the journal of the Association for Psychological Science (APS), “the highest ranked empirical journal in psychology” (Sage Publications, n.d., n.p.), which has authors adhere to the APA Style Manual (2010), as of January 1, 2014, emphasizes in its submission guidelines to authors: “to include effect sizes accompanied by 95% confidence intervals rather than standard deviations or standard errors” (APS, n.d.).
The APS goes further by describing the need for published research to be transparent in terms of reporting sufficient detail for research replication. Furthermore, data analysis emphasis needs to shift from NHST to reporting confidence intervals (CI) and effect size (ES) in an effort to demonstrate the practical importance of findings. Finally, suggestions are emerging in this community for (a) the establishment of dissemination venues for replication studies and (b) commitments from researchers to their various research ethics boards that they will report all research analyses, not just those that result in “significant difference” (APS, n.d.).

Research transparency. Very typically, research journals impose word limits on the length of articles they will publish. This forces authors of these articles to make decisions about where to “prune” words in order to stay within the word limits. Several journals, with Psychological Science leading the way, have moved to not including word counts from the methods or results sections:

Effective January 2014, the Method and Results sections of either Research Articles or Research Reports do not count toward the total word limit. The aim here is to allow authors to provide clear, complete, self-contained descriptions of their studies, which cannot be done with restrictions on those sections. (APS, 2015, n.d.)

This shift is not to encourage authors to simply become more verbose in their descriptions, the emphasis is on providing fulsome, but cogent, descriptions of the methods and results. Anyone reading a research article written in this journal should be able to replicate, identically, the research being reported. Educational policy makers and practitioners should look for complete descriptions of research methods and results in any research they read. The method section of any research report should provide the step-by-step recipe for how the study was conducted and how the data were analysed. When descriptions are lacking, the conclusions and recommendations should be treated with caution since it is impossible, as a consumer of that research, to know what was omitted and how that might affect the results being reported.

Confidence intervals and effect size. The discussion of the practical importance of findings is one that should be of interest to educational policy makers and practitioners, alike. It is not unusual to find studies using NHST approaches reporting “significant” findings that are practically unimportant. Since significance in NHST analyses is a function of, both, effect size and sample size, a study having a large sample size, even when the observed effects are small, is very likely to lead to the conclusion that a significant finding has been identified. Two main approaches for addressing this issue have been identified in the literature: CI and ES.

Reporting of CIs provides consumers of research with a sense of the precision of any estimated result. For example, pollsters regularly survey people on a variety of
topics, these are sometimes shared via news or other media. While descriptions of the methods used are very terse, usually the pollsters will report how many people were surveyed (the “n”), the percentage who responded in each particular way (the “point estimate”), and the amount of potential error present in the point estimate (the estimate of the possible range the point estimate actually represents): the confidence interval. Using the example of the pollster who might report on the proportion of the population who supports a particular political party in an upcoming election, a point estimate and confidence interval might be reported as 35%, with a margin of error of plus or minus 5%, of people surveyed support Party X. Given the size of the sample the consumer knows that as little as 30% of the actual population might support Party X or as much as 40% of the population might hold this opinion. This is not a new concept and extensive descriptions of confidence intervals for parametric and non-parametric measures of association and difference are available (e.g., Glass & Hopkins, 1996; Hopkins et al., 1995; Pedhazur, 1997; Sprinthall, 2011).

The second approach, reporting ES, addresses the issue of practical importance also from the point of view of the problem created by studies making use of large sample sizes. In such studies, minor differences between groups or weak correlations between variables can result in significant findings given NHST approaches. These trivial group differences or weak correlations are of very little value in educational settings. Imagine for a moment that a study finds that PISA scores can be raised by 0.1% across an entire province or country by implementing a particular instructional strategy; furthermore, this strategy will only require an increase to the provincial or national education budget of 5%. Most people would agree that the cost would not justify the gain: the finding, while statistically significant, is not important. A budget increase of this magnitude should be expected to result in a greater substantive gain. ES addresses this issue by taking into account the gain without the effect sample size plays on NHST. Put another way, ES provides the reader with a sense of the magnitude of the relationship between variables or the differences between groups.

Just as with determining CIs, methods for determining ES for parametric and non-parametric measures of association and difference have existed for many decades (e.g., Glass & Hopkins, 1996; Hopkins et al., 1995; Pedhazur, 1997; Sprinthall, 2011). Examples of ES measures of association include correlation ($r$), explained variance ($r^2$), coefficient of contingency ($C$), Cliff’s $d$, Cohen’s $d$ (probably the most common ES indicator used in parametric tests of difference in social sciences research), Hedges’s $g$, eta-squared, omega-squared, and many others. While the purpose of this chapter is not to provide in depth understanding of these measures, it may be useful to understand how to interpret ES coefficients.

It is noteworthy that there is not universal agreement as to the exact break-points for small, medium, and large effect sizes. Approximate values of small, medium, and large effects for three sample ES measures are shown in Table 1 (Sprinthall, 2011).

Using ES to gauge the magnitude of differences between groups or the magnitude of the association between variables provides consumers of research with a metric
that does not change conceptually from study to study and does not suffer from the
pitfalls of NHST approaches.

In addition to looking for and using practical approaches to establish the utility
and magnitude of effects, using measures of association or of difference, educational
policy makers and practitioners are also urged, as suggested by the APS (n.d.) and
Cumming (2013) to look for cumulative evidence. Synthesizing the findings across
research studies in a form of meta-analysis (Glass, McGraw, & Smith, 1981) is
beyond the ability, and often, interest of educational policy makers and practitioners.
Consumers of research are encouraged to seek out meta-analytic research that
systematically brings together the works of different researchers studying the same
or very similar research questions. A variety of frameworks for this type of synthesis
have been proposed, a few include: (a) Meta-Analysis Reporting Standards as
proposed by the APA (2010), (b) Cooper (2010) provides a meta-analytic framework
consisting of seven inter-related steps, and (c) Cummings (2013) provides a brief
description of techniques for combining CI results from different studies.

While research consumers who lack the expertise to conduct these statistically
based meta-analyses may feel overwhelmed by the suggestions above regarding
meta-analysis of research findings, this is not to say that a statistical approaches
are the only way to synthesize research findings. Less formal approaches drawing
on qualitative analysis techniques can also be employed which thematically cluster
studies’ contributions on particular topics. This is particularly true if careful attention
is paid to understanding the meanings of CI and ES when reported.

RESEARCH AND THE CONTEXT OF POLICY AND DECISION MAKING

Some of the early writing around policy, specifically the adoption of innovations
from an adult education perspective, would lead the reader to believe that policy
development and implementation is a relatively linear process in which credible
evidence simply needs to be obtained (i.e., through research) and disseminated (e.g.,
Rogers’ diffusion of innovation theory—Rogers, 1962, 2003) to the appropriate
people for implementation. Building on the work of Bryce Ryan and Neal Gross
(1943), who drew on the work of Gabriel Tarde (1903), the French sociologist
responsible for describing the “S” shaped diffusion and adoption of innovation curve
with Iowa farmers and their adoption times for new hybrid corn (i.e., innovators,
early adopters, early majority, late majority, and laggards), Rogers (1962)
differentiated between diffusion and adoption in that diffusion was a process that occurred among groups of people as they learned about new approaches. Adoption, he argued, was an individual process in which a person made a commitment to implementation. To reach implementation commitment, Rogers identified five stages individuals move through: (a) first developing knowledge of the innovation, (b) developing an attitude toward the innovation, (c) making a decision to adopt or reject the innovation, (d) decide how to implement the innovation if it is seen as worthwhile, and (e) confirm that the decision was a good one. Rogers also identified four pre-conditions impacting whether a decision is made to implement an innovation or not, these include: (a) the decision maker’s previous practices, (b) whether the individual felt the innovation potentially addressed a need or problem, (c) the innovativeness of the decision maker, and (d) norms of the social context within which the individual operated. Finally, of critical relevance to the central discussion in the present paper was the role of mass media; Rogers argued that an individual’s awareness and understanding of an innovative idea is most effectively influenced through the use of mass media and its ability to inform, in lay terms, innovators of leading research and innovations.

Looking for policy development literature reveals a perceived gap between research and practice exists, surprisingly, in the field of medicine. In the first of a series of eight papers focused on the gap between research and practice in medicine, Haines and Donald (1998) argued that

...how best to promote the uptake of research findings has been fuelled by a number of factors including the well documented disparities between clinical practice and research evidence of effective interventions. Examples include interventions in the management of cardiac failure, secondary prevention of heart disease, atrial fibrillation, menorrhagia, and pregnancy and childbirth. (p. 72)

Haines and Donald (1998) went on to argue that “there is also growing awareness that conventional continuing education activities, such as conferences and courses, which focus largely on the passive acquisition of knowledge have little impact on the behaviour of health professionals” (p. 73). Drawing on Rogers’ (1983) diffusion of innovation work, Haines and Donald crafted their argument about how to close the research practice gap by focusing on the care provider and the patient. Keeping in mind that Haines and Donald’s target audience was health care professionals, they identify two general categories of barriers preventing the implementation of research in their practice: environmental and personal. Within these two categories they identified specific areas of barriers; those which influence or are influenced by the government level policy makers in the health care area include:

1. Lack of financial resources,
2. Health policies which promote ineffective or unproved activities,
3. Influence of the media on patients in creating demands or beliefs,
4. Impact of disadvantage on patients’ access to care,
5. Obsolete knowledge,
6. Influence of opinion leaders, beliefs and attitudes (e.g., a previous adverse experience of innovation),
7. [Patient] demands for care, and
8. [Patient] perceptions or cultural beliefs about appropriate care

The parallels to education are strikingly similar!

In another British study focusing on an historical view of what has driven higher education policy in the U.K., Shattock (2006) argued that from the 1930s to present, policy decisions have not been driven by research conducted by scholars within the higher education community. Shattock examined the tension between the internal Higher Education political forces (the “inside”) which exist within the Higher Education institutions themselves and those political forces from various levels of government who often control purse-strings (the “outside”). He notes that prior to the mid-1970s, in the U.K., higher education policy was driven primarily from the inside, but since that time, policy has been driven from the outside. Shattock cited the 2003 White Paper, *The Future of Higher Education* (DfES), in which six themes driving U.K. higher education into the future were affirmed; these he demonstrates are derived “from a Treasury belief, fostered by OECD and the European Union…” (p. 136).

An examination of the American post-secondary education system suggests that higher education in that country is driven from the “inside” by the market forces that directly affect the institutions (Trow, 2003). Interestingly, Trow’s assessment of the American higher education policy shaping forces paralleled Shattock’s assessment of the U.K. situation with respect to the lack of influence by research conducted by scholars focused on higher education—research appears to be irrelevant.

There are a number of lessons this brief sampling of policy development from three fields suggests for broad scale policy development and decision making in basic education. First, if we are to believe the innovation diffusion theory perspective, the key is to identify innovators (i.e., risk takers and able to cope with uncertainty) and early adopters (i.e., people who have a high degree of “opinion leadership” and who serve as role models in their communities) to promote ideas (popular media is one important means) to stakeholder groups (i.e., educators, parents and business community members—the voters, various levels of government). Indeed these innovators and early adopters are likely members of the stakeholder groups just mentioned. Haines and Donald’s (1998) contribution to medical policy making helps educators to identify similar barriers that must be overcome in order to implement new educational approaches; these are:

1. Lack of financial resources;
2. Educational policies promoting ineffective or unproved activities;
3. Influence of the media on parents and students in creating demands or beliefs;
4. Impact of disadvantage on students’ access to education;
5. Obsolete knowledge;
6. Influence of opinion leaders, beliefs and attitudes (e.g., a previous adverse experience of innovation);
7. Parent and student demands for education; and
8. Parent and student perceptions or cultural beliefs about appropriate education.

The insight that is potentially most troubling in this exploration comes from the higher education field. Clearly, policy making in basic education shares much with post-secondary education. The higher education experience in the U.K. and the U.S. would suggest that research from within the academy is not particularly relevant; not because such research does not exist, nor because it does not have something relevant to say about higher education, but because politicians and other decision makers, whether they are from the “inside” (i.e., university presidents) or the “outside” (i.e., government ministers) typically choose not to pay attention to it. They are more focused on pre-existing ideology and beliefs than they are on what scholarship from the field has to offer.

Innovation in k-12 policy making and decision making generally follows the same patterns described by Rogers (1962). In particular, educator decision making is very much dependent on Rogers’ pre-conditions for adoption of innovative thinking or practice. Typically, evidence contradicting existing beliefs requires suspension of those beliefs rather than simply ignoring the contradictory evidence because it does not fit with decision-maker pre-dispositions. This requires explicit effort to accomplish as, often, contradictory evidence is simply discounted without a thorough analysis of its veracity.

BRINGING THE THREADS TOGETHER

This chapter calls for changes in how research is viewed, critiqued, and used by educational policy makers and practitioners. Returning to the ideas underpinning this chapter, that is, the need for educational policy makers and practitioners to draw on research findings for enacting changes in policy and practice, a variety of observations are offered.

First, field research is extremely messy because researchers are unable to control for all of the unexpected possible threats to validity. Having said that, field research also has great potential for contributing to our understandings of educational processes since teaching and learning takes place in field settings. It does not take place in carefully controlled laboratories.

Consequently—the second point that needs to be highlighted—consumers of research must develop fundamental understandings of research design and be acutely aware of the potential threats to the different types of validity so they can assess the robustness of the research method. In other words, consumers of research
need to be able, with confidence, to identify threats to the validity of different types of research, which can cause rival explanations for results to become tenable. This requires understanding research method to identify where threats to validity may exist as well as understanding educational context to be able to imagine the competing rival explanations for research findings. In field-based research there is no perfect research method, anything a researcher does to mitigate one validity related issue invariably creates or has the potential to create a methodological problem: minimizing the threats to validity while maximizing the trustworthiness of research becomes the quest.

Third, consumers of research need to develop, minimally, a level of understanding of statistical data analysis techniques that enable them to comprehend the ideas underlying those techniques. Knowing some fundamental approaches and their theoretical underpinnings allow for research consumers to understand the basic elements of research focused on exploring hypotheses of difference and research focused on exploring hypotheses of association. This also means understanding the importance of not violating the assumptions underlying the statistical approaches, or, alternately, recognizing when statistical assumptions have been violated and what that means in terms of the research findings being reported.

Fourth, consumers of research are encouraged to seek research that de-emphasizes NHST approaches, instead looking for research that reports the results of quantitative analyses in very practical terms of CI and ES. The educational policy maker and practitioner is reminded to be very skeptical of research that focuses on and emphasizes NHST approaches discussing findings on the basis of low p-values only.

The final point that will be highlighted is the importance of drawing on the insights gained from multiple studies. This is not an exercise in finding only those studies that resonate with the educational policy maker or practitioner. Without a question, it is possible to find research that supports virtually any point of view; however, if our purpose is to advance practice on the basis of research, it is critical that the literature considered on any topic be balanced and evaluated through a critical lens.

Addressing these issues will enable educational policy makers and practitioners to make changes supporting educational improvement. To do so, though, requires that decisions be made based on credible evidence gathered from multiple studies that are selected to provide balanced views of the state of the art. Both educational policy makers and practitioners need to be aware of their biases and beliefs, hold them in suspension as advocated by Dewey (1933), to avoid drawing only on research aligned with their pre-dispositions for informing their decisions and actions.

NOTE

1 One-tailed and two-tailed tests of significance and the effect of these on alpha levels will not be dealt with here, for in depth explanations of this, please refer to any introductory statistics text.
REFERENCES


