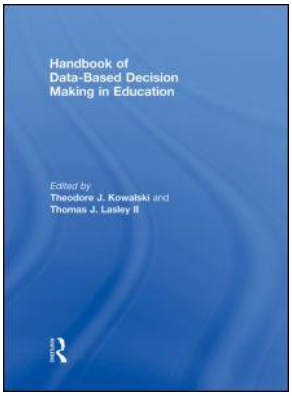
This article was downloaded by: *10.3.97.143*
On: 29 Nov 2023
Access details: *subscription number*
Publisher: *Routledge*
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: 5 Howick Place, London SW1P 1WG, UK



Handbook of Data-Based Decision Making in Education

Theodore J. Kowalski, Thomas J. Lasley

Research and Evaluation on Data-Based Decisions in Education

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9780203888803.ch14
Gregory J. Marchant, Sharon E. Paulson
Published online on: 13 Oct 2008

How to cite :- Gregory J. Marchant, Sharon E. Paulson. 13 Oct 2008, *Research and Evaluation on Data-Based Decisions in Education from:* Handbook of Data-Based Decision Making in Education Routledge
Accessed on: 29 Nov 2023
https://www.routledgehandbooks.com/doi/10.4324/9780203888803.ch14

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Handbook of Data-Based Decision Making in Education

Edited by
Theodore J. Kowalski and
Thomas J. Lasley II

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

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

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

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

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

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

© 2009 Taylor & Francis

All rights reserved. No part of this book may be reprinted or reproduced or utilized in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data
Handbook of data-based decision making in education / Theodore J. Kowalski & Thomas J. Lasley II, editors.
p. cm.
Includes bibliographic references and index.
1. School management and organization–Decision making–Handbooks, manuals, etc. I. Kowalski, Theodor 1943– II. Lasley II, Thomas J. 1947–
LB2805 .H2862 2008
371.2 22

ISBN 0-203-88880-4 Master e-book ISBN

ISBN10: 0–415–96503–9 (hbk)
ISBN10: 0–415–96504–7 (pbk)
ISBN10: 0–203–88880–4 (ebk)

ISBN13: 978–0–415–96503–3 (hbk)
ISBN13: 978–0–415–96504–0 (pbk)
ISBN13: 978–0–203–88880–3 (ebk)

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

14

Research and Evaluation on Data-Based Decisions in Education

Gregory J. Marchant and Sharon E. Paulson
Ball State University

The purpose of this chapter is to identify the characteristics of good research that can be applied to educational decision making and the subsequent evaluations of those decisions. For years, research has been evaluation's uppity brother, having strict rules and requirements for asking and answering empirical questions in educational settings; whereas evaluation could be as informal as asking, "Well, how did you like it?" In turn, researchers and evaluators might argue that data-based decision making is even less rigorous than evaluation. Although important educational decisions are being made using data-based evidence, the results of the decisions are inconclusive unless research principles are used in making the decision and empirical evaluations are conducted to see if the desired outcomes have been attained. As these educational decisions and evaluations turn increasingly to objective data and research methodologies, they begin looking more like their rigid brother. In this chapter, we argue that, given the right tools, data-based decision making can be as reliable and valid as good evaluation and research. In fact, using the right tools and approaches, data-based decision making is indeed an alternative form of evaluation and research.

Data-Based Decision Making and Educational Evaluation

Data-based decision making has grown out of the need to eliminate guesswork in developing curriculum, creating programs, or changing instructional practice. Educational policies, including those driven by No Child Left Behind (NCLB), now mandate that data be collected in almost every educational arena to determine the success of schools, teachers, and students. Schools and school districts have built massive databases of test scores, drop-out rates, school attendance, disciplinary actions, and financial expenditures. By using these data, teachers and administrators can make decisions regarding curriculum, instruction, and programming that will improve the success of their schools and their students. There are a number of definitive sources on how to use data to make decisions (e.g., Kowalski, Lasley, & Mahoney, 2008), including this book; it is not the purpose of this chapter to repeat those methods. However, both the data and the decisions must be derived using sound research principles. In turn, once a decision is made, one needs to determine whether or not the outcome of interest has been realized. Data-based decision

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

making cannot stop with the decision (and its implementation); the decision must then be evaluated to ensure its effectiveness. Indeed, those who have written extensively on data-based decision making have argued that the basis for educational decisions must be grounded in educational evaluation (Kowalski et al., 2008).

Driven in the 1960s by policy requirements to evaluate educational programs, educational evaluation initially meant *measurement*, not unlike what is done today to produce the very educational data of which we speak. Popham (1993) argued, however, that measurement is *status appraisal*, whereas evaluation is *quality appraisal*. Evaluation goes beyond simply collecting the data (measurement) and includes appraising their worth. *Evaluation is the process of delineating, obtaining, and providing useful information for judging decision alternatives* (Stufflebeam et al., 1971). Over the past 20 or more years, numerous methods and models of educational evaluation have emerged, but ultimately the goal of all evaluation is to determine whether or not some measurable outcome was attained. Although evaluation can be as simple as asking, "Did the participants like it?" (where *it* is curriculum, instruction, program, or other decisional process), the most common form of evaluation currently used to assess educational practice is *outcome-based evaluation* that answers the question, "Did it work?" (Popham, 1993; Schalock, 2001). More recently, *theory-based evaluation* which answers the question, "Why did it work (or not work)?" has been added to the repertoire of evaluation techniques in education (Rogers, Hacsi, Petrosino, & Huebner, 2000; Weiss, 1997). Both of these types of evaluation require strict research principles including asking the right questions, planning a valid research design using reliable and valid measurement, analyzing the results statistically, and drawing appropriate conclusions.

In the sections to follow, we will explore the concepts and methods of conducting empirical research to serve as a lens for viewing data-based decision making in education. Topics include asking the right questions, reviewing the existing literature, developing a purpose to guide methodology, using valid procedures and measurement, analyzing results, and drawing appropriate conclusions. Educators should be mindful of these techniques at all points of the decision-making process: when collecting the data, when using the data to make decisions, and when evaluating the decisional outcomes. To demonstrate these research principles, we will use several examples of data-based decisions that we have encountered in our own research and evaluation efforts.

The Question and its Purpose

For years there has been an emphasis on the importance of the research question to any study. It is the question that specifies the topic of the study, defines the outcomes of interest, and directs the research methods. Obviously, these are key elements in any data-based decision as well. For example, school district A is concerned with high school dropout rates and with pass rates on the state standardized achievement test. It is likely that these data have been collected and the district wants to use these data to make decisions; in particular, to decide what can be done to reduce drop-out rates and to increase pass rates on achievement tests. Similarly, teacher B has concerns over

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

her students' grades: what can I do to help my students get higher grades (with the assumption that higher grades is a reflection of greater learning)? Neither district A nor teacher B is going to randomly make changes to curriculum and instruction to see whether drop-out rates, pass rates, or grades change over time. To address these concerns, the school district or the teacher needs to formulate specific, well-crafted questions. Principles of research propose several types of research questions that might be used to address such concerns (Meltzoff, 1998).

- *Questions of description*: These questions address the characteristics of the data. For example, what are the characteristics of students who drop out of school; or what are the qualities of students who make higher grades?
- *Relationship questions*: These questions address whether or not a relationship exists between two factors. Do pass rates on standardized achievement tests relate to drop-out rates? Is student attendance related to grades?
- *Causality questions*: These questions address the factors that influence the outcome of interest. For example, does the school mentoring program affect drop-out rates; or does peer-tutoring increase students' test scores?
- *Interaction questions*: These questions take into account the specific circumstances under which an outcome or relations among factors exist. Does the school mentoring program lower drop-out rates more for students with higher grades than for those with lower grades?

Formulating the specific question(s) to be asked requires a great deal of insight into the issues related to the question(s) and the context of the data. Research articles in education start out with a literature review for just this reason, to establish the background (Boote & Beile, 2005). Likewise, those in decision-making positions need to know the literature related to the questions they are formulating. Despite what some might think, some pretty definitive research results already exist concerning many educational policies and practices. For example, take the question, "Does retaining students who do not pass the state standardized test improve their long-term achievement?" The research on retention has already determined that the answer to that question is "no," decision made (Jimerson & Kaufman, 2003). "Does retaining students who do not pass the state standardized test increase drop-out rates?" That answer is "yes"; again, decision made (Jimerson & Kaufman, 2003). We already know that student retention is not good practice, so making the decision to utilize it would be poor policy. Already knowing that, a school district might focus its questions on other interventions that could be developed and tested for students not passing *the* test.

In addition to acquiring a knowledge base concerning research findings and the educational context, decision makers need to establish the ultimate purpose behind their question(s). Good research creates consistency among the questions to be addressed, the purpose of the questions, and the methods used to collect the data (Newman, Ridenour, Newman, & DeMarco, 2003). The purpose behind a decision is of paramount concern; otherwise the whole data-based process could be completed without informing any actual decision.

Different types of purposes have been identified for research that would apply to

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

data-based decision making (Newman et al., 2003). For example, data are often collected to examine a school or district's power or organizational structure, priorities, practices, or policies. In addition, school data might be used to measure outcomes or the consequences of certain practices. Data can be used to examine the past, or to generate or test new ideas. Sometimes data are used to understand complex situations and/or to add to the knowledge base. Data can also be used to predict events based on past information, or to inform constituents about present conditions. The key is to know why the data are being collected or considered, so that what is being asked of the data is appropriate.

Figure 14.1 is commonly presented to consider the issue "Which line is longer, AB or CD, or are they the same?" A similar but much more complex question is often posed in educational contexts: Which program is more effective? Program A, Program B, or are they equally effective? There are at least two ways to answer both of these questions. One way is to look at the perceptions and opinions of those involved, and the other is to use some formal, objective method of measurement.

So, what is the purpose of the data to be collected? Is it to be used to present the differences in people's perceptions of the length of the lines, or is some actual measure of length required? About half of the people asked believe that line CD is longer than line AB with most others thinking they are the same length. When measured, the lines are in fact the same length. The first purpose, to discover people's perceptions, might provide insight into human beliefs; whereas the second purpose is to find the actual relation between the lengths of the two lines: two different questions, two different purposes, two different answers. Similarly, questions asked by decision makers in education often involve issues of perception versus some other outcome or measurement. Is the concern whether constituents value or like a policy or program, or whether that policy results in some other measurable outcome? For example, surveys generally find that parents would like to be able to choose the school their children attend; however, there is little evidence that school choice results in higher achievement (Lubienski, 2007; Zimmer et al., 2007).

Now take, for example, the teacher trying to decide how to increase students' grades in the classroom. The teacher with two sections of the same course wanted to test whether or not reciprocal learning (students teaching each other) is more effective than lecture. To do so, he alternated methods between the two classes twice and gave a test after each lesson. He also gave a survey asking which method the students preferred. The results indicated that the achievement was the same across the two

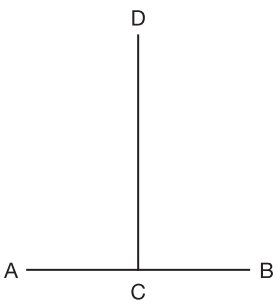


Figure 14.1

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

teaching methods, but the students greatly preferred the reciprocal learning approach. Being an advocate for reciprocal learning, the teacher was disappointed that the reciprocal learning groups did not achieve more than the lectured groups. He thought that the *purpose* of his study was to establish the superiority of the more innovative approach based on achievement. If that were the case, he should have involved a number of instructors and only collected test results. However, because he was the only instructor involved, had the reciprocal learning groups scored higher, it would only have meant that the students could teach each other better than *he* could. Although the instructor considered the student preference data secondary and almost inconsequential compared to the achievement data, collecting the data meant there was another purpose to his study. He wanted to know what the students preferred. Therefore, within the context of the limited scope of just his instruction, he learned that he could use a teaching approach that he advocated and that his students preferred with no decrease in achievement.

Validity and Getting at the Truth

The second major research concept that is important to any data-based decision is validity. In paradigms of logic, validity is the *truth* of a statement. In research, validity is the most important aspect of the study. In a sense, it is the *truth* of the study. If the results are valid, they represent the truth, and not some alternative explanation. There are those who believe that it is possible to get data and statistics to support any position available. However, there is an inherent truth in any data set and the results gleaned from it. That truth may provide a clear direction to guide decision making, or that truth may be obscured or inconclusive because of flaws in the research design. These flaws are threats to the validity of research and distractions from the truth. Many of these problems have been spelled out in the literature on research methods (Cook & Campbell, 1979).

Selection

By far the biggest threat to the truth in education is the selection of the participants included in the data. In true experimental research, participants are randomly selected from a larger population and randomly assigned to treatment groups. The idea is that any differences inherent in the participants will not be over- or under-represented in the sample or any subgroups (those getting or not getting a treatment or intervention). However, random selection and assignment are almost nonexistent in real educational environments.

The samples of students and teachers that researchers get to study are often the same sets available to administrators (or less so). These samples are not random nor are they comparable. For example, we have found that more than 50% of the differences among elementary schools in pass rates on standardized tests can be attributed to the race and income of the students (Marchant, Ordonez, & Paulson, 2008). Any effort to attribute school differences to special programs or to faculty must be

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

examined in light of the demographic characteristics of the students. Similarly, when the College Board releases the list of SAT test scores averaged for each state, newspapers report front-page stories indicating the rise or fall in the rankings of their states suggesting that this has something to do with the quality of a state's schools. Unfortunately, the differences among states have little to do with the quality of the states' educational programs. The truth is that over 90% of the differences among states can be tied to two variables: the percent of Black test takers in each state and more importantly the education level of the test takers' parents (Marchant & Paulson, 2001). The overall percent of high school students taking the test for each state really does not matter (although the College Board claims this is a major factor). Therefore, state averaged SAT scores do not provide valid data for making decisions regarding effective state-level educational policies because of the selection differences among states.

Selection problems can occur at any level: state, district, school, or classroom. At the classroom level, most students are randomly assigned to teachers at each grade level. However, boys at an urban elementary school might need "a positive male role model," so these boys, who also tend to be lower achieving, are assigned to the few male teachers in the building. A review of annual achievement data finds the male teachers' classes are not achieving as high as other classes in the school. Are these male teachers less effective? Of course not, but when students with demographic characteristics known to be associated with lower test scores are disproportionately present in a classroom, school, district, or state, it follows that average scores at those levels will be affected.

The word "volunteer" associated with any data should be a red flag. Simply put, those who volunteer are often different from those who do not. Therefore, any comparisons between volunteers and non-volunteers are likely to reflect this difference, not any program the volunteers are involved in. Due to instructional freedom, teachers' unions, and simple courtesy, teachers usually are not forced to try certain programs or practices. It is not unusual for a publisher or a researcher to be introduced at a faculty meeting looking for volunteers to try a particular program or approach. Often that program requires time for training and possibly classroom observations of the teachers who volunteer. The data at the end of the study show that the teachers who volunteered, who were probably more dedicated and more confident, had students with higher achievement; but did the special program have anything to do with the achievement? We simply cannot know.

Maturation

Another confounding event, particularly in education data, is maturation of the students. A few years ago we were involved in a meta-evaluation of a popular Title I math program for an urban school district (Paulson & Marchant, 2001; Paulson, Marchant, & Rothlisberg, 1999). The program employed special instructors to teach elementary students algebra using group involvement techniques. The program was received positively by both students (there were no tests) and teachers (they had an hour a day off). An evaluation of the program (which doubled students' math time)

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

showed positive effects on state achievement tests in math. However, looking at the achievement broken down by grade level provided some insight. Our review of the program found little or no advantage for students at the lower grade levels; however, when students reached fifth grade and the program's algebra content matched the curriculum and was developmentally appropriate for the students, the extra math time mattered and the students excelled. Therefore, the achievement seemed less a result of the special nature of the costly program, but more a function of time-on-task as the students matured.

Testing

Our national obsession with testing requires close attention to a couple of problems associated with testing. The simple act of taking a test can improve a student's performance on subsequent tests; familiarity with the content and format of tests is likely to help a student on future tests regardless of instruction or intervention. Any special program that includes the use of a test similar to the outcome measure is likely to yield higher performance regardless of the nature of the rest of the program. Therefore, any gains in performance attributed to No Child Left Behind that mandate annual testing must be tempered with the fact that scores should increase even if "real learning" does not.

Mortality

We conducted a study, exploring whether having a high school graduation exam was related to states' SAT scores, that could have fallen victim to two threats to its validity: increased testing and mortality (Marchant & Paulson, 2005). Mortality occurs when members of a data set leave or they are removed from the data in a non-random fashion (mortality is almost never random), and this change in the data then influences the outcome of the study. In this study, we speculated that high schools that required a graduation exam would have a curriculum that is more focused on the exam content and less involved in the type of reasoning skills measured by the SAT, subsequently resulting in lower SAT scores. However, we were concerned that the states that required a high school graduation exam would have students who were more familiar with testing, thereby effectively increasing their SAT scores. In addition, states with graduation exams have lower graduation rates (Amrein & Berliner, 2002; Jacob, 2001; Marchant & Paulson, 2005), so lower performing students would not be taking the SAT; again, effectively raising a state's average SAT scores. Consequently, even if the presence of graduation exams had a negative impact on reasoning skills, the threats to validity would diminish such findings. Surprisingly, our analyses found that the graduation exams were related to lower SAT scores despite the confounding effects of testing and mortality. Had we not been aware of these possible threats, however, we may not have found the truth.

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

Instrumentation

Another threat to a study's validity involves measurement. In the process of collecting data, any change in the instrument (e.g., version of a survey or test, criteria or cutoff scores, or method of record keeping) may result in changes in scores over time that could be incorrectly attributed to another cause (e.g., student learning or quality of teaching). As state departments of education began to understand the requirements and procedures for demonstrating adequate yearly progress (AYP) for No Child Left Behind, they realized that the percentage of students passing the state achievement test was the functional criterion for avoiding negative consequences. Therefore, some states worked to lower their cutoff scores (Cavanagh, 2005). Simply looking at changes in pass rates in these states would suggest that learning had increased because more students were now passing the test, perhaps because of the implementation of NCLB. Obviously, such a conclusion would not be valid.

On the classroom level, teachers also collect data to inform their instruction. Increasingly, teachers are using pre-tests before their instruction to determine what students have learned from a lesson. To determine what students have learned from a pre-test to a post-test, the two instruments need to be similar: comparable in both content and difficulty. However, teachers sometimes make their pre-test relatively easy because the students have not been taught the material; then the post-test covers the more difficult new material. It is quite possible that students will do better on the easier pre-test than the more difficult post-test. Someone analyzing the data is left with the conclusion that the students became more ignorant because of instruction, when in reality no comparison can be made because of the threat to validity called instrumentation.

Regression to the Mean

Regression to the mean may help a remedial reading teacher (personal experience), but it can cause problems when teachers do not understand its threat to the validity of their classroom data. There is an established trend in testing that individual scores tend to gravitate toward the mean of a group of scores; so the students who score inordinately high on the first administration of a test are likely to score lower (closer to the mean) on a subsequent testing. Similarly, the students who score very low on a reading test are likely to have improved scores on the next test (score closer to the mean) even if the instruction was not particularly effective. Although this effect is usually small, it can be significant and create problems for data interpretation. Gifted programs that appear to be stagnant in achievement may actually be performing quite well, just as remedial programs may display false improvement. In large data sets with normal distributions, these effects tend to offset each other over time; therefore, although local school district scores may go up or down slightly, the overall scores for the state remain relatively stable.

In sum, research has identified many of the potential flaws in data that can get in the way of accurate analyses and well-informed decisions. There is a poignant saying in statistical analysis of data, "Garbage in, garbage out." If there is a problem with the

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

data before they are analyzed, the results and conclusions in turn will be flawed. Using data at any level, classroom, school, or statewide, to make potentially important educational decisions requires a keen awareness of these potential threats.

Results, Analyses, and Statistics

Research is designed to describe, to look for relations among factors, or to look for differences among groups. The same is true for data-based decisions. Sometimes what is needed is just a deeper understanding of a situation; sometimes there is a need to know whether two or more events are related in some way, maybe such that some outcome can be predicted (like drop-outs); other times the issue concerns determining differences among two or more groups, perhaps between those using a special program and those who are not. In any case, once you know the purpose of the study, you can formulate specific questions that can guide the data you collect, how you collect them, and finally how you analyze them.

Analyzing data is probably one of the most difficult aspects of good research-based decision making. Often teachers, administrators, or other educators do not have the training to conduct appropriate statistical analyses on their data. For this reason, they produce simple statistics like means or frequencies, and such analyses rarely represent much by themselves. Data-based decision makers should seek the expertise to conduct meaningful analyses. This is the one aspect of data-based decision making that might require a good relationship with a local college or university. Academics in higher education can either provide the training or the consultation that will make any data-based decisions more effective. Although this section is not meant to provide the information that one would need to analyze data effectively, it will describe the techniques and terminology used in statistical analyses that can inform decision makers, whether they are merely reading the literature or deciding what types of analyses might be appropriate for the questions and purposes being examined.

Questions of Description

Descriptive research does just that (and only that); it describes what happened, without seeking to establish a cause or to determine group differences. It simply describes through qualitative narratives or quantitative means and distributions a unique situation or an unanticipated event. Statistics in a descriptive study are usually not very complicated: means and standard deviations. A mean is the arithmetic average: the scores (any quantitative measure) added up and divided by the number of scores. A standard deviation is the average distance from the mean for all of the scores. If most of the scores are close to the mean, the standard deviation is small. If there are large differences among the scores so that the range is great and the scores are spread out, the standard deviation is large. In some situations the distribution and standard deviation are as important as the mean in understanding data. For example, a group of students have a low mean and a small standard deviation on a test. An intervention is employed that raises the mean, but it greatly increases the

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

standard deviation. Such results suggest that the intervention really helped some students a lot, but did not help others nearly as much. If you looked at the mean without considering the variation, you might assume that all students improved because of the intervention; a conclusion that might lead to less than adequate decisions in the future. Furthermore, to provide meaningful information for decision making, more complex questions and analyses are required.

Relationship Questions

To answer questions regarding relations among factors, correlation analyses are most common. A simple relationship can be addressed using a simple correlation (known as Pearson *r*); but questions addressing the relations among multiple factors would require multiple regression analyses. A multiple regression in simple terms is just a multiple correlation; the statistical analysis establishes how several factors are related to one outcome. To revisit some of our earlier questions: what factors are related to higher standardized test scores or lower drop-out rates? A multiple regression would report whether the factors are related to the outcome and which factors are more important to the relationship. It is crucial to remember, however, that correlation does not mean causation. To actually test whether or not an outcome was caused by some program or intervention, a more controlled design is required. In the case of most educational decisions that might be made at the classroom, school, or district level (testing a curriculum or program, for example), the most common way to answer the question of causation is to assess change over time or to compare groups (e.g., the group with the intervention to one without).

Causation Questions Using Change Scores or Group Differences

To show change, some type of pre-test or pre-measure is needed. As mentioned, the pre-measure needs to be the same in content and difficulty as the post-measure. Identical measures can be problematic due to familiarity to the specific instrument. The change between a pre-measure and a post-measure can be statistically tested to assess whether or not the change was large enough to be significant (i.e., that it was not due simply to chance). The simple statistic used to test the difference between two sets of scores is a *t*-test. Simply subtracting the pre-measure from the post-measure yields a change score that can be used in this analysis.

Similar to change scores, these same types of analyses can be used to analyze differences among groups. When an intervention or treatment (e.g., instructional technique or program) is examined for a group, that group is referred to as the experimental or treatment group. The group that does not receive the treatment is called the control group. Having a control group that is comparable to the treatment group is critical to determining the effectiveness of any treatment. Unfortunately, obtaining control groups can be a politically tricky endeavor. If there is a special population eligible for a program, it just makes sense to place all of these students (or the most needy of the students) in the program. However, if all of the eligible

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

students receive the intervention, then comparisons on a range of outcomes including achievement between very needy students and not so needy students are likely to underestimate the effectiveness of the program. For example, a group of students, who are behind one grade level and progressing at about half a grade level each year, are placed in a special program where they progress three fourths of a grade level in a year. Such a change in achievement may represent substantial improvement for these students and indicate a successful program. Unfortunately, threats to validity, like selection and maturation, make it difficult to know whether the program itself was responsible for the change (maybe the students just matured during the course of the intervention and made greater gains because of normative changes in skill). A comparable control group who did not receive the treatment would show whether or not the change occurs naturally, in which case it would occur in the control group too, or whether the change occurred because of the intervention, in which case the control group would not show improvement over the same period of time. However, if this needy treatment group with .75 grade-level improvement is compared to a control group of regular classroom children who demonstrate their normal growth of one whole grade-level achievement, one might incorrectly conclude that the needy students would have achieved more had they not received the special program. Fortunately, there are statistical techniques to control for these preexisting differences between experimental and control groups, but they are never as good as having a control group formed by randomly assigning students to either the treatment group or the control group. In educational settings, however, random assignment is usually neither politically nor ethically likely.

The more sophisticated analysis of variance (ANOVA) is needed to make comparisons among multiple groups (e.g., does school climate differ across four elementary schools in a given school district?) or to make comparisons among multiple factors that might differ between groups (e.g., does school climate differ by grade level and by school?). The ANOVA is similar to running multiple *t*-tests (just like a multiple regression is similar to running multiple correlations). The advantage of ANOVAs (and a troubling and confusing thing as well) is that the results for the group interactions are known as well. In this way, an administrator knows which schools and which grade levels reported a better school climate (these are known as main effects), but the administrator may also learn that fourth and fifth grades at two of the four schools have a significantly better school climate than the rest of the classes. This may provide much better direction for decisions than either of the main effects alone.

Interaction Questions

Most decisions in education are complicated, such that simple answers are either not enough or simply wrong. Interactions exist among most factors that might be measured in educational settings. It is not unusual for analyses to be conducted without concern for interactions (this is called a Type 6 Error; Newman, Marchant, & Ridenour, 1993). In the earlier ANOVA example regarding school climate, it is possible that both or neither main effect (grade level and school) was significant. Only knowing the main effects might have suggested that school climate was always better

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

at upper-grade levels across all schools or that school climate was always better at two particular schools across all grade levels; or it could have indicated that none of the grade levels or schools were better than the others. Examination of the interaction would show that only two grades at two schools had a higher school climate. Perhaps the factor that is affecting school climate in two of the elementary schools is that the schools have designated content specialist teachers for the upper two grade levels. Knowing the true nature of the differences among schools and grade levels will lead to more accurate decisions for intervention or programming.

Another benefit of ANOVA, that is shared with multiple regression, is the ability to include covariates, usually to statistically control for differences among groups. The function of a covariate is to partition out the effect of a confounding variable that might threaten the validity of the study, so that other factors may be examined more accurately. In our study that explored the relations between graduation exams and SAT scores (Marchant & Paulson, 2005), one of the research questions was: Do states that require graduation exams have lower SAT scores? Going into the study, there was a confounding problem because over 90% of the differences among states have been associated with the education level of the test takers' parents and the percent of minority test takers in the states. Because these characteristics are not randomly distributed among states, it is possible that demographic variables like these could influence the results. Therefore, when the statistical analyses were performed, demographic variables were included as covariates to control for pre-existing differences among states. Although this approach is by far inferior to random assignment, it does reduce the effect that the known factors can have on the outcome.

Finally, it is important to mention that although good data are more important than more data, more data are good, and generally the more times something is measured the better. Multiple measurements before an intervention help to determine whether a trend existed that carries on through the treatment. In other words, what appeared to be an effect of an intervention could merely be the continuation of a trend. More important is the continuation of any effect after intervention. After an inservice, teachers feel very positively about a particular practice, but have they increased that practice a week later? If so, are they still using the particular practice a year later? There has been a great deal of research on Head Start and similar programs with most of it reaching the conclusion that immediate effects of these programs are decidedly positive; however, if there is no continued support for these students, the benefits diminish and disappear (Barnett & Hustedt, 2005). Therefore, regardless of the design and analysis, the purpose of the study should suggest the need for short-term and/or long-term results.

Interpreting and Misinterpreting Results

In conducting sound research, there is one final principle: drawing accurate conclusions. Any data-based decision that has heeded the preceding principles of research and evaluation will yield accurate conclusions from the results of whatever questions were being addressed. Even if those making decisions are not collecting the data or conducting the research, it is important that they know what they are looking for and

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

not accept results blindly. Knowing what to look for in data and research allows decisions to be made correctly, or as importantly, provides insight into when to reject the data or the conclusions others have drawn.

Earlier the concept of validity was described in the context of arguments in logic as the truth. Assuming there were no serious factual or procedural errors, there are still threats to the truth. The context in which data are presented can inappropriately influence their acceptance, either positively or negatively. Not all data or research are from reliable sources, but not all questionable sources should be dismissed. Bias, real and perceived, is a threat to correctly identifying valid information for decision making.

Schools, school districts, and the state and federal governments collect much of their education data by involving academics from local colleges and universities. Moreover, many of the analyses of educational data are run in academic settings, although the federal government does some of its own research, and state departments of education issue many of their own reports. Most of what local schools and governmental agencies report are descriptive studies. Test scores or other indicators of achievement may be reported as going up or down, but there is seldom any type of sophisticated statistical analyses of what the data mean. This information is often released as reports and/or released to the media. Unfortunately, there are no reviews of the reports or their conclusions before the results are disseminated. Therefore the truth of what is presented is usually the interpretation of the results by the media. Test scores or other indicators change slightly (oftentimes due to chance) and suddenly public concerns are raised over the quality of education in a particular school, district, or state.

In contrast, research completed in colleges and universities (often funded by government agencies or private foundations) is usually presented at research conferences and published in professional journals. Before being accepted for presentation or publication, the studies must undergo blind review (authors and institutions are not identified) to establish the quality of the research, and other researchers critique the study to identify validity problems. Although professional organizations may accept half or more of the studies they review for their conferences, some journals of empirical research publish fewer than 5% of the manuscripts they receive for review. Any research that has undergone the scrutiny of academics or other research professionals is more likely to be a valid indicator of the state of educational practice.

One of the concerns of reviewers of empirical research is bias: Did the researchers ignore threats to validity and/or come to conclusions that likely misrepresent the truth because they were guided by a desire to obtain a particular result? The review process is an important step in safeguarding the public and other professionals from this type of misinformation. In contrast, when reports are directly issued and/or released to the press, they sidestep the review process. Fortunately, school and government reports that are more descriptive do not tend to present causal inferences or analyses that are open to much misinterpretation. Media sources are more likely to be guilty of presenting interpretations that inaccurately reflect what is being reported.

More recently, there has been a proliferation of think-tanks aligned with conservative policy agendas that issue research reports directly to the media. Although the think-tanks claim to be nonpartisan, their funding sources are not, and the research is

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

often critical of public education practices and supportive of privatization efforts like vouchers and charter schools. This research has usually not withstood the review process of other empirical research endeavors, so their findings might need to be considered in that light. In an effort to provide academically sound reviews of this type of research, the Think-Tank Review Project (http://epsl.asu.edu/epru/thinktankreview.htm) was created as a collaboration between the Education Policy Research Unit at Arizona State University and the Education and the Public Interest Center at the University of Colorado. The reviews of this research have identified problems that range from slight errors in interpretation to fatal flaws in research methods and gross misstatements of facts. The Project also annually presents The Bunkum Awards that "highlight nonsensical, confusing, and disingenuous education reports produced by think tanks" that "have most egregiously undermined informed discussion and sound policy making" (Education Policy Research Unit, 2006). For example, the 2006 runner-up for the award "collected data, analyzed [the data], and then presented conclusions that their own data and analyses flatly contradicted" (Welner & Molnar, 2007).

Our review of the Title I math program previously mentioned found that student maturation influenced the results. This thorough evaluation of an existing evaluation is called a meta-evaluation. For this particular program quite a few evaluations had been conducted to assess the program's effectiveness, most by a former school administrator who became associated with the project. All of these evaluations were very positive about the impact of the project on math and even English achievement (that in itself raised questions for us: why would a program specific to algebra improve English skills?). There were only a couple of evaluations available from other sources, both of which were negative and resulted in the school districts involved dropping the program. The conclusion of our meta-evaluation of the program in a large midwestern urban school district was inconclusive in that the data presented in the evaluations did not clearly substantiate their positive conclusions (Paulson, Marchant, & Rothlisberg 1999).

Consumers of research and educational decision makers need to be careful not to blindly accept or reject data-based information. This is true of publicized accounts in the media, as well as more local reporting of data. Fallacies can undermine the correct interpretation of data and lead to misinterpretations that cause either incorrect acceptance or rejection of information. Fallacies are errors in reasoning that, if better understood, are easier to avoid (see Table 14.1, adapted from Labossiere, 1995). As with any information, decision makers need to consider their sources. However, they need to be knowledgeable enough of research principles to judge the data, research, and conclusions independent of the sources.

Future Directions

In the past, the types of research questions addressed by academic researchers and school administrators have been somewhat different. Academia was interested in investigating psychological concepts, theoretical models of learning, and general teaching approaches (like cooperative learning). School districts, on the other hand,

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

Table 14.1 Thirteen fallacies and examples that can detract from data for decision making.

Fallacy	*Example statement*
Characteristics of the person	"Of course you found our students do better in reading than math, you're an English teacher."
Previous position	"During our last negotiations you thought the health care provisions were fine, so how can you say there is a problem now?"
Position of authority	"I don't care about the problems you claim to have found with the program, Dr. Smith says it leads to more achievement."
Appeal to a belief	"Everyone knows that the schools in this country are failing, so they must be bad."
Common practice/tradition	"We have always used basal readers, why are you trying to change things?"
Appeal to emotion	"The students and the teachers all love Project C, so it must be doing great things."
Appeal to novelty	"This new math curriculum is fresh off the press, it is going to be great."
Appeal to ridicule	"Those who can't do, teach; now they even want more money for supplies, right!"
Burden of proof	"You have not proven that cooperative learning works, so we're going back to lecturing."
Confusing cause and effect	"We tend to have lower expectations for poor low achieving students. So, those poor students do not achieve as much because of the low expectations."
Hasty generalization	"Ms. Artbunkle's class didn't benefit from the extra tutoring, so why waste money on it for the other classes?"
Misleading vividness	"Of course violent crime is up in schools, don't you remember hearing about the awful shooting in that school last month?"
Straw man	"You want to reduce emphasis on standardized testing. Fine, let's just eliminate all accountability for schools."

wanted information about their specific schools and specific programs. Although there was some overlap between the two areas of inquiry, schools were put off by much of the statistics and research jargon. Similarly, researchers wanted answers to the big questions and did not like being limited to a context that would make it difficult to generalize their findings.

Recently there has been a convergence of perspectives. Among decision makers in education, there is a growing acceptance of methods once considered the domain of educational researchers, as educators are realizing that getting to the truth is the important next step to having data support their decisions. Approaches once considered obstacles, such as control groups and statistical analysis, are now accepted as important tools for making sound decisions; and being able to demonstrate results in an objective, justifiable way is now the currency of funding and programmatic decisions. Likewise, educational researchers are acknowledging the importance of context and recognizing the function of exploring specific programs that operationalize their

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

concepts (Ball & Forzani, 2007). Although much theoretical research still exists, studies regarding bigger concepts in carefully controlled laboratory settings are very few. The bottom line now for most research is: Can this study improve school learning or inform educational policy? And the answer to that question does not stop with statistical significance, either.

Even the statistics presented in educational research have become more applied. The most popular statistical test of significance, based on the number of participants in the study and variability on the measures, indicates the likelihood that relations among variables or differences among groups is due to chance. With a large sample size, very small differences or relationships could be statistically significant, but are they relatively important? For example, in a study of effects of high stakes testing on state math achievement (Marchant, Paulson, & Shunk, 2006), both demographic differences among students and high stakes testing indicators across states were significant indicators of state math achievement. However, the demographic differences across states accounted for 78% of the variance, whereas differences in high stakes testing indicators accounted for only 3% (NAEP Math). Both were statistically significant, but relatively speaking, how important were the high stakes testing indicators?

Consequently, the statistic now used more commonly to reflect the relative impact of a factor is called an "effect size." One indicator of effect size uses a standardized form of means and standard deviations to assess the relative difference between two groups on some outcome. In this case, the mean for the control group is set at zero with a standard deviation of one (similar to *z*-scores) and the effect size tells where the mean for the experimental group would be on that scale if the factor of interest were present (e.g., some program or teaching method). In a popular article from 1984, Bloom reviewed the effect sizes generated from all of the research on effective teaching from the 1970s and 1980s in an effort to find methods of group instruction as effective as one-to-one instruction. One-to-one instruction had been found to have an effect size of two. This meant that compared to traditional instruction, an equivalent group of students receiving one-to-one instruction would achieve with a mean two standard deviations above the mean of the group receiving traditional instruction (this is a huge difference). This translates to about 98% of the one-to-one instruction group scoring above the mean of the traditional instruction group (welcome to Lake Wobegon where all the kids are above average).

Effect size is an extremely useful tool for those making data-based decisions, especially in conducting cost-benefit analyses in schools. In education, costs and benefits seldom use the same unit of measure; whereas costs can usually be measured in money or time, benefits are often a measure like student achievement or graduation rate. Effect size allows one to conceptualize the size of a benefit. If there are few costs associated with a programmatic benefit in school achievement, the decision is easy. If there is a major cost and little benefit determined by a small effect size, the decision is also easy. It is when costs and benefits are both considerable that decisions become more difficult. Although effect size does not put costs and benefits on the same scale, it is an effort by educational researchers to give decision makers a common gauge for considering outcomes.

The discussion of cost-benefit analysis points out the difficulty that educators face

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

in data-based decision making that does not exist in the business world. This is important to elucidate because there are always those trying to impose a "business model" on education and educational decision making. As in medicine, the ultimate outcome of the process of education results in the quality of life of people. It is difficult to answer questions like: How much is a program worth that is likely to keep 10 students a year from dropping out? A school could calculate the funding it loses from a student dropping out, but that is not the point of education. There is a societal and humanitarian imperative that goes beyond dollars and cents. How much is it worth for a disadvantaged child to achieve a complete grade level in one year versus something less? How much is that worth to the child? To the parents? To society?

The decisions faced by educators are some of the most difficult of any profession. The outcomes are sometimes abstract, long-term, and difficult to gauge. Simple information may be useful, but it can also be wrong. Although the immediate concern must be the education of students, teachers and parents and the general public all have a vested interest in the education process. Turning guesswork into data-based decision making is a step in the right direction, while finding the truth in the data and the results by adopting educational research procedures is another step. Decision making in education will never be a perfect science or be able to function under a strict business model but making the best decision possible has always been the goal, and data bases and research methodology are changing what is possible in a very positive way.

References

Amrein, A. L., & Berliner, D. C. (2002). *An analysis of some unintended and negative consequences of high stakes testing* (Research document). Tempe, AZ: Education Policy Research Unit, Education Policy Studies Laboratory. Retrieved January 15, 2008, from http://epsl.asu.edu/epru/documents/EPSL-0211–125-EPRU.pdf

Ball, D. L., & Forzani, F. M. (2007). What makes educational research "educational?" *Educational Researcher*, *36*, 529–540.

Barnett, W. S., & Hustedt, J. T. (2005). Head Start's lasting benefits. *Infants and Young Children*, *18*, 16–24.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16.

Boote, D. N., & Beile, P. (2005). Scholars before researchers: On the centrality of the dissertation literature review in research preparation. *Educational Researcher*, *34*, 3–15.

Cavanagh, S. (2005). Illinois board lowers bar on English-learners' test. *Education Week*, *24*, 25.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation*. Boston: Houghton-Mifflin.

Education Policy Research Unit (2006). *Bunkum awards in education*. Retrieved October 4, 2007, from the Arizona State University Education Policy Studies Laboratory website: http://espl.asu.edu/epru/epru_2006_bunkum.htm

Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis*, *23*(2), 99–121.

Jimerson, S. R., & Kaufman, A. M. (2003). Reading, writing, and retention: A primer on grade retention research. *The Reading Teacher*, *56*, 622–635.

Kowalski, T. J., Lasley, T. J., & Mahoney, J. W. (2008). *Data-driven decisions and school leadership: Best practices for school improvement*. Boston: Pearson.

Labossiere, M. C. (1995). *Fallacy tutorial pro 3.0, A Macintosh tutorial*. Retrieved December 25, 2007, from The Nizkor Project website: http://www.nizkor.org/features/fallacies/

Lubienski, C. (2007). Charter schools, academic achievement and NCLB. *Journal of School Choice*, *1*(3) 55–62.

Marchant, G. J., & Paulson, S. E. (2001). State comparisons of SAT Scores: Who's your test taker? *NASSP Bulletin*, *85*(627), 62–74.

Marchant, G. J., & Paulson, S. E. (2005). The relationship of high school graduation exams to graduation rates and SAT scores. *Education Policy Analysis Archives*, *13*(6). Retrieved October 26, 2007, from http://epaa.asu.edu/epaa/v13n6/

Marchant, G. J., Ordonez, O. M. M., & Paulson, S. E. (2008, August). *Contributions of race, income, and cognitive ability to school level achievement.* Paper to be presented at the American Psychological Association, Boston.

Marchant, G. J., Paulson, S. E., & Shunk, A. (2006). Relationships between high-stakes testing policies and student achievement after controlling for demographic factors in aggregated data. *Educational Policy Analysis Archives, 14*(30). Retrieved October 26, 2007, from http://epaa.asu.edu/epaa/v14n30/

Meltzoff, J. (1998). *Critical thinking about research: Psychology and related fields.* Washington, DC: American Psychological Association.

Newman, I., Marchant, G. J., & Ridenour, T. (1993, April). *Type VI errors in path analysis: Testing for interactions.* Paper presented at the American Educational Research Association, Altanta, GA (ERIC Document Reproduction Service No. ED 362 529).

Newman, I., Ridenour, C., Newman, C., & DeMarco, G., Jr. (2003). A typology of research purposes and its relationship to mixed methods. In A. Tashakkori & C. Teddie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 167–188). Thousand Oaks, CA: Sage.

Paulson, S. E., & Marchant, G. J. (2001, April). *Does Project SEED improve math achievement and self-concept?* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Paulson, S. E., Marchant, G. J., & Rothlisberg, B. A. (1999). *Review of Project SEED* (unpublished evaluation). Muncie, IN: Educational Program Evaluation Center, Ball State University.

Popham, W. J. (1993). *Educational evaluation* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Rogers, P. J., Hacsi, T. A., Petrosino, A., & Huebner, T. A. (Eds.) (2000). Program theory evaluation: Practice, promise, and problems. *New Directions for Evaluation, 87*, 5–13.

Schalock, R. L. (2001). *Outcome-based evaluation* (2nd ed.). New York: Kluwer.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Hammond, L. R., Merriman, H. O., & Provus, M. M. (1971). *Educational evaluation and decision-making in education.* Itasca, IL: Peacock.

Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. *New Directions for Evaluation, 76*, 41–53.

Welner, K. G., & Molnar, A. (2007, February 28). Truthiness in education [Commentary]. *Education Week, 26*(5). Retrieved January 2, 2008, from http://epsl.asu.edu/epru/articles/EPSL-0702–407-OWI.pdf

Zimmer, R., Gill, B., Razquin, P., Booker, K., Lockwood, J. R., Vernez, G. et al. (2007). *State and local implementation of the No Child Left Behind Act: Volume I—Title I school choice, supplemental educational services, and student achievement*, Retrieved December 15, 2007, from the Department of Education website: http://www.ed.gov/rschstat/eval/choice/implementation/achievementanalysis.doc

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14

Downloaded By: 10.3.97.143 At: 04:28 29 Nov 2023; For: 9780203888803, chapter14, 10.4324/9780203888803.ch14