Strengthening Inferences in Quantitative Education Studies Conducted by Novice Researchers: Capitalizing on Standards for Sampling, Research Design, and Instrumentation

Mohammed A. A. Abulela
South Valley University, Egypt / University of Minnesota, USA

Michael Harwell
University of Minnesota, USA

Introduction

The goal of many quantitative education studies is to produce valid and replicable findings that add to our knowledge and understanding in ways that improve subjects' outcomes, for example, identifying the most effective way to teach number ratios (e.g., $\frac{3}{7}$) to increase student learning in middle school mathematics classes. All methodological components of a study are important, but we focus on sampling, research design, and instrumentation because of their central role in the validity and replicability of study findings (internal and external validity), because recommended standards for these components do not appear to have received the attention they deserve, and because researchers may find these particularly challenging. In addition, the data analysis component is a broad field that requires many articles to cover the recommended standards.

In planning a study, researchers can turn to several resources offering guidance in the form of recommended methodological standards such as the Social Science Research website (http://www.socialresearchmethods.net/kb/), the American Psychological Association (APA) Publications and Communications Board Working Group on Journal Article Reporting Standards (2008), and What Works Clearinghouse [WWC] (2017). Research methodology texts (e.g., Pedhazur & Schmelkin, 1991), as well as summaries of research methodologies (U.S. Department of Education, 2013; Ellis & Levy, 2009, 2010), also offer useful resources. These resources provide information for capitalizing on these standards in planning and executing a study that enhances the likelihood of valid and replicable study-based inferences. However, these standards presuppose a level of expertise and experience that may not be present among novice researchers such as new faculty, individuals beginning non-faculty roles such as a working in a university-affiliated research center or a government-funded education center, faculty transitioning to more research-oriented work, and students conducting their own research.

The goal of this paper is to encourage novice researchers in the educational sciences to capitalize on methodological standards and to respond to methodological challenges that can undermine these standards in ways that help preserve the validity and replicability of study-based inferences. We begin by reviewing recommended methodological standards for sampling, research design, and instrumentation and provide examples of common methodological challenges and advice on how to respond. We also review a sample of quantitative studies to assess the extent to which the standards have been employed.

Review of Methodological Standards

We assume a study's rationale, literature review, and research questions appropriately inform the quantitative methodology, and that ethical guidelines like those endorsed by the American Psychological Association (see https://www.apa.org/ethics/code/)) have been observed.

Sampling

Sampling focuses on the way subjects (e.g., consumers, households, students) are obtained for inclusion in a sample. How subjects were sampled speaks directly to the generalizability of study findings, which depends on specifying a population to generalize study findings to. Sampling can also involve study conditions, settings, instruments, etc. (Shadish, Cook, & Campbell, 2002), but we focus on subjects such as students and schools. By definition, a population is a collection of subjects that cannot usually be accessed which we wish to generalize study findings to, for example, all seventh grade mathematics students in the upper Midwest of the U.S. in the 2018-2019 school year. A sample is a chunk of a population that is used to generalize to a population of interest. The importance of generalizing study findings cannot be over-emphasized: A fundamental goal of much quantitative research is to identify interventions or conditions that improve outcomes like mathematics achievement for large numbers of subjects. The extent to which study findings can be accurately generalized to a population is often referred to as external validity (Campbell & Stanley, 1963), i.e., how externally valid are study findings? The way a sample was obtained speaks directly to external validity.

Types of Sampling Mechanisms

Two basic sampling mechanisms are available: probability-based (random) sampling and non-probability (non-random) sampling. The former is the recommended standard (WWC, 2017) because it supports strong external validity whereas the latter typically does not.

In probability-based sampling, a population is specified, a sampling frame consisting of a list of every subject in the population is constructed, each subject in the population is assigned a unique identifying number (id), and a random process is used to generate a sample of N numbers. The latter usually involves a random numbers program in computer software like SPSS (IBM Corp., 2011) or Stata (Stata Corp, 2015). In the final step, N subjects with ids corresponding to the random numbers produced by the software are selected for the sample. This process ensures the probability of sampling each subject in a population is known (i.e., $\frac{1}{N_{pop}}$ assuming a large population and a comparatively small sample, N_{pop} = number of subjects in the population). The logic underlying generalizability is simple: Because each subject has the same opportunity to be sampled, characteristics of a sample (e.g., gender, socio-economic status or SES, mathematics achievement) should mimic those in a population and hence sample results should be generalizable to the population.

Common probability-based sampling mechanisms include simple random sampling, stratified random sampling, and cluster (two-stage) sampling (Lohr, 2010). Suppose a researcher wishes to generalize study findings of the impact of a new mathematics curriculum for teaching number ratios to a population of seventh grade students in mathematics classes in the upper Midwest of the U.S. in the 2018-2019 school year. If N_{pop} = 20,000 and probability-based sampling is used, a unique id is assigned to each student in the population via the sampling frame that might range from 1 to 20,000. To obtain a sample of, for example, N = 1,200 a computer program could be used to produce 1,200 random numbers between 1 and 20,000 to identify N students to be sampled. The size of the sample is often a function of statistical power in the data analysis and the choice of N = 1,200 in this example is arbitrary. A simple random sample of size N consists of N subjects selected from the population such that every set of N subjects has the same probability of being sampled.

Stratified random sampling involves carving a population into strata such as those based on SES (e.g., high, medium, low) with known percentages of subjects in each stratum (e.g., 20%, 60%, 20%), and then using simple random sampling to obtain the desired number of subjects from each stratum. Stratified random sampling is used when it is important that the composition of the population (e.g., high/medium/low SES) be represented in the sample, which is not assured if simple random sampling is used. In cluster (two-stage) random sampling a population of organizational units (clusters) such as schools is specified, J clusters are sampled at random (stage 1) and in the typical case all subjects within a cluster are sampled (stage 2).

Despite the strong external validity associated with probability-based sampling, it appears to be uncommon in most educational studies, almost certainly because of the resources needed for probability-based sampling. For example, surveys of published educational studies provide evidence of high rates of non-probability-based sampling, such as Dedrick et al. (2009) and Fath (2014) who reported that 73% of 99 surveyed studies and 93% of 58 surveyed studies used non-probability-based sampling. Probability-based sampling is typically used in large studies such as the U.S.-based Early Longitudinal Childhood Studies (ELCS), or international studies such as Trends in International Mathematics and Science Study (TIMSS, 2015) or the Program for International Student Assessment (PISA, 2017).

Practical constraints associated with probability-based sampling often result in studies using non-probability-based sampling involving regional or locally-obtained samples. For example, it may not be possible to generate a sampling frame because the number of subjects in a population is unknown, meaning the probability of sampling each subject in a population is unknown. The most common form of non-probability-based sampling occurs when a sample is obtained because it is convenient, which often takes the form of sampling all available subjects locally (e.g., all students in an introductory psychology class in a university) (Battalgia, 2008). Accurately generalizing study findings based on a convenience sample is challenging because of uncertainty about the population the sample represents, leading to weak external validity.

Still, it is important to try to generalize study findings from non-probability-based sampling. An appropriate strategy is for a researcher to provide empirical evidence the convenience sample is similar to a population of interest on selected indicators. This advice is consistent with the report of Wilkinson and the APA Task Force on Statistical Inference (1999):

“Using a convenience sample does not automatically disqualify a study from publication, but it harms your objectivity to try to conceal this by implying that you used a random sample. Sometimes the case for the representativeness of a convenience sample can be strengthened by explicit comparison of sample characteristics with those of a defined population across a wide range of variables.” (p. 595)

Suppose a convenience sample of N = 300 students was obtained from three schools in an upper Midwest state but the study goal was to generalize to all seventh grade mathematics students in the state. Providing information about key indicators for the sample and the population of interest can strengthen the case for generalizability. For example, evidence that the percentage of urban/suburban/rural students, Black, Hispanic, and White students, and students in poverty in the sample is similar to that in the state provides indirect evidence supporting generalizability; similarity of educational indicators for the state and the sampled schools such as average pupil/teacher ratios, per pupil spending, and percentage of students who attend a post-secondary institution provides additional evidence of generalizability (In the U.S. the latter information is available through the Digest of Education Statistics and the Bureau of the Census).

Two other data characteristics can provide important challenges to valid generalizations. One is attrition (missing data) which occurs for a variety of reasons. For example, once a sample is identified students may initially participate in a study but then move, parents may refuse to allow their children to participate in a study or to complete a study, or schools that initially agreed to participate subsequently decide not to. These kinds of sampling difficulties lead to missing data and can seriously undermine the generalizable of study findings because sampled students, schools, etc., with missing data may differ from those who provided complete data in ways that impact inferences.

A second challenge to valid generalizations is lack of independence of subjects and hence their data, which can seriously distort inferences particularly from data analyses. Both probability-based and non-probability-based sampling can produce dependency but the latter is typically more likely, for example, all students in a school are sampled (convenience sample) which may include siblings, students studying together in a class, etc. Researchers should carefully examine the sampling process for evidence of dependency, which, if present, can have a devastating impact on inferences.

Examples of common methodological challenges

As suggested above, a common methodological challenge in sampling is attrition, which can undermine generalizability. We provide a brief example of this methodological challenge for each sampling method and advice for novice researchers.

In simple random sampling, suppose a researcher receives funding to examine the impact of a new online method for teaching English to non-native English speakers in the seventh grade as measured by a test of written English. The research question is: Does online teaching of English to seventh grade non-native speakers improve their written English? The population of interest is all seventh grade non-native English speakers in the upper Midwest of the U.S. who have indicated that Somali, Hmong, Spanish, or French was their native language. The researcher decides a simple random sample of N = 600 non-native English speakers is needed. Schools and students already using the online method for teaching English are excluded.

Suppose 20,000 eligible students define the sampling frame based on information provided by the U.S. Department of Education. Parental consent is required for a student to participate in the study, and the researcher anticipates 50% of the students will not be eligible to participate because their parents will not consent based on previously reported rates in the literature (which represents attrition). After assigning each student in the population a unique 5-digit id ranging from 00001 to 20,000 the random numbers option in SPSS (IBM Corp., 2011) is used to generate a sample of 1,200 5-digit values between 00001 and 20,000 and those students are selected to be in the sample. A letter is then sent to the parents of the 1,200 students asking them to allow their student to participate in the study--if 50% consent the resulting sample will be (.50)(1,200) = 600. The consent rate turns out to be 55% meaning (1,200)(.55) = 660 students are eligible. To help ensure generalizability of study findings to the desired population it is important to provide evidence the final sample of 660 students is similar to the initial sample of 1,200 students. This evidence could take the form of comparing information for students in the population and sample (e.g., socio-economic status, race, gender, test scores), and information about the schools in the population and sample (e.g., percentage of Black, Hispanic, and White students, school poverty rate).

In stratified random sampling, using the information provided in the above example, the researcher decides that the proportion of native speakers of Somali, Hmong, Spanish, and French in the population of non-native English speakers should be represented in the sample and chooses stratified random sampling. Suppose the proportions of Somali, Hmong, Spanish, and French speakers in the upper Midwest are known to be 20%, 20%, 55%, and 5%, respectively. The four strata in this example consist of (20,000)(.20) = 4,000, (20,000)(.20) = 4,000, (20,000)(.55) = 11,000, and (20,000)(.05) = 1,000 students. Assuming a consent rate of 55% (N = 660), applying simple random sampling within strata involves sampling (660)(.20) = 132 Somali and 132 Hmong native speakers, (660)(.55) = 363 native Spanish speakers, and (660)(.05) = 33 native French speakers (N = 660). To help ensure generalizability of study findings to the desired population it is important to provide evidence the final sample of 660 students is similar to the initial sample of 1,200 students. Using stratified random sampling enables the researcher to appropriately generalize study findings based on N = 660 to students in the four strata.

In cluster (two-stage) random sampling, using the information provided in the first example, a researcher is interested in investigating characteristics of teachers and their likely impact on the written English of non-native speakers in the upper Midwest. The researcher decides a random sample of J = 40 teachers (clusters) is needed and expects 30% of sampled teachers to decline to participate in the study (which represents attrition). Suppose J_{pop} = 700 teachers and N_{pop} = 20,000 eligible non-native speakers of English. A sampling frame of the 700 eligible teachers is generated in which each teacher has a unique 3-digit id. The random numbers option in SPSS (IBM Corp., 2011) is used to generate a sample of J = 60 teachers (to take the expected attrition into account) (stage 1) and their students (N = 1,800 assuming 30 students per class, stage 2) are selected to be in the sample. Assuming the actual consent rate is 70% produces 60 - (1-.70)(60) = 42 teachers and (42)(30) = 1,260 students (The latter will shrink if student consent rates are less than 100%). Empirical evidence that the sample of 42 teachers and their students are similar to the original sample of J = 60 and N = 1,800 helps to ensure generalizability of study findings to the desired populations.

In convenience sampling, using the information provided in the first example, the researcher decides to use a convenience sample of N = 1,200 non-native English speakers in the seventh grade. The researcher goes to three nearby schools who have agreed to participate in the study and, assuming the consent rate is 55%, produces a sample of 660 students. Demonstrating this sample is similar to the population of seventh grade students in the upper Midwest is critical to generalizability arguments.

Practical advice

We offer five pieces of advice on sampling to novice researchers assuming a quantitative study in education is to be performed:

1. The importance of external validity argues for probability-based sampling but this sampling method is uncommon. Consider using a large publicly available dataset like Early Longitudinal Childhood Studies in the U.S. and TIMSS or PISA internationally because these studies employ probability-based sampling. Datasets of this kind may also contain collateral information that provides evidence the resulting sample is similar to the population of interest on key indicators such as poverty rates.

2. The sampling process should be carefully examined to help ensure subjects are independent because dependency in the data subjects provide (e.g., mathematics test scores) can have a devastating effect on inferences from data analyses that assume independence such as multiple regression. Evidence of dependency should prompt actions to help ensure independence of subjects, for example, if siblings are identified randomly select one sibling for inclusion in the sample and omit the remaining sibling(s).

3. If non-probability-based sampling is used, it is particularly important to provide empirical evidence supporting generalizability, such as the similarity of student demographics in the sample and the population of interest, as well as similarity of important school-based indicators like per pupil spending or poverty rates. Similar summaries support at least some generalizability, but if there is little similarity between the empirical evidence for a sample and the population of interest external validity will be severely compromised and the value of the study undermined. Researchers should plan on collecting relevant information to make these empirical comparisons.

4. In practice, more than N subjects would often be sampled because the final N will likely be smaller due to attrition. In all instances, provide evidence that subjects who attrited are similar to those who participated in a study.

5. It is important to report inclusion and exclusion criteria of the sampling procedure. For example, restrictions on age, race, or socio-economic status (SES) of the sample or the use of strata should be made clear (Appelbaum et al., 2018). This information will help clarify the population which study results ideally can be generalized to.

Research Design

Research design represents the

"... glue that holds the research project together. A design is used to structure the research, to show how all of the major parts of the research project—the samples or groups, measures, treatments or programs, and methods of assignment—work together to try to address the central research questions." (Trochim & Land, 1982, p. 1)

A study's research design is important because it largely dictates the strength of causal inferences that can be drawn from study findings. Studies that credibly draw strong causal inferences answer the question that permeates quantitative research: Does A cause B and if so under what conditions and for whom? Harwell (2011) encouraged researchers to select a research design supporting strong causal inferences, and the Council for Exceptional Children (2014) similarly stated that "causality could be reasonably inferred from research designs when they are well designed and conducted" (p. 1). Summarizing the strengths and weaknesses of different designs based on methodological standards and pointing novice researchers towards stronger designs should promote stronger inferences. Note that providing evidence of a causal effect requires a design with at least two conditions (typically groups), often labeled treatment and control, which constitute an independent variable.

A critical feature of causal inferences is that estimates of a treatment effect are unbiased, for example, the difference between the outcome (dependent variable) means of treatment and control groups represents, beyond random variation, the causal effects of an independent variable. Put another way, strong causal inferences imply there are no serious alternative explanations for the observed effect of an independent variable on an outcome, implying strong internal validity (Shadish, Cook, & Campbell, 2002). A study with weak internal validity implies weak external validity even if probabilistic sampling was used because internally invalid results cannot typically be validly generalized.

Types of Research Designs

What Works Clearinghouse (2017) describes three categories of group-based research designs to choose among: Randomized control trials (RCTs), quasi-experimental designs (QEDs), and regression discontinuity designs (RDDs).

Randomization is often characterized as the gold standard of research design because its produces probabilistically unbiased treatment effect estimates (i.e., if the study was repeated many times, the average treatment effect will be unbiased), supporting strong causal inferences. This occurs because randomization ensures (probabilistically) that subjects in the treatment and control groups are equal in observable and non-observable characteristics at the study onset (e.g., SES, gender, mathematics achievement), and consequently differences between groups after the treatment has been applied and its effects measured are attributed to the treatment and not other (confounding) variables (e.g., SES, gender). Traditionally, randomization is described as assigning N subjects at random to a treatment or control group using a random process (there can be multiple treatment groups), such as a coin flip or generating N random binary numbers using a computer program such that subjects receiving a 1 are assigned to treatment and those receiving a 0 to control. Thus, the probability of a subject being assigned to the treatment or control group is .50.

By definition, a control condition implies the treatment under study is not present and a true control group is one in which subjects receive no treatment of any kind related to the study. For example, if an RCT was used to study the effects of different doses of caffeine (e.g., 0 milligrams, 250mg, 500mg, 1000mg) on cognitive functioning the 0mg condition represents a true control group because these subjects receive no caffeine. Evidence of a difference in treatment-control outcome means would (beyond random variation) be attributed to the causal effect of caffeine. However, control conditions often represent “business as usual” in that subjects receive the current treatment or practice. For example, seventh grade students in a control group continue to use the current mathematics curriculum in a school whereas those in a treatment group use a new curriculum, or nurses in a control group in a hospital interact with patients in traditional (“business as usual”) ways (about three minutes per interaction), whereas those in a treatment group limit their interactions with patients to about one minute. Thus “business-as-usual” resembles a treatment and a difference between treatment and control means under random assignment is likely due to the treatment, but the possibility remains the business-as-usual curriculum strengthened or weakened the treatment-control mean difference.

Despite the critical advantage of unbiased treatment effects, practical and policy constraints on assigning subjects to treatment or control groups often make RCTs impractical. For example, students assigned at random to a treatment group may need to be pulled out of class to participate in a study which a teacher or parent may not allow, asked to remain after school which schools or parents may balk at, or students randomly assigned to a control group are subsequently moved to the treatment group at the insistence of a parent. A popular version of random assignment that responds to many practical and policy constraints of traditional RCTs involves assigning higher level units (clusters) such as schools at random to treatment or control conditions, and are known as randomized cluster designs. The growth of hierarchical data analyses (Raudenbush & Bryk, 2002) has played a key role in the popularity of randomized cluster designs (Kleinman, 2017). Hierarchical (multilevel) analyses analyze data obtained from hierarchical structures in which lower level units, such as students, are clustered within higher level units (clusters), such as schools. Hierarchical data in these analyses are the result of two-stage (cluster) sampling.

For example, suppose the impact of a new curriculum for teaching number ratios to seventh grade students is studied by randomly assigning J = 42 schools to either a treatment condition in which teachers in a school use the new curriculum (treatment), or a control condition in which teachers use the existing (business-as-usual) curriculum. All students complete a test of number ratios at the conclusion of the study which serves as the outcome variable. The fact that all students in a school are in the treatment or control condition is likely to make random assignment more acceptable while still producing unbiased estimates of the treatment effect. Appelbaum et al. (2018) encouraged researchers to describe the units of randomization as well as the procedures used to generate the random assignment sequence, for example, assigning classrooms using randomly generated 0s and 1s.

Practical or ethical constraints on random assignment often leads researchers to employ QEDs. Quasi–experimental designs are used to compare pre-existing groups which define the independent or “treatment” variable. In some cases, the pre-existing conditions cannot be randomly assigned such as SES status (high, medium, low) or gender, whereas in other cases an outcome for subjects already receiving a treatment, such as mathematics achievement scores for students participating in an existing mathematics curriculum, are compared against those of students participating in a new curriculum. The problem with QEDs is that the lack of random assignment means group differences on an outcome may be due to a treatment effect, pre-existing differences between treatment and control groups that affect the outcome (selection bias), or both. Put another way, the probability a subject is assigned to a treatment or control group is not .50 meaning groups are not probabilistically equal and the likelihood of biased estimates of a treatment effect may be high.

For example, treatment schools (teach a new curriculum) may have higher SES than control schools (use an existing curriculum), which leads to the former having higher outcome means because higher SES is usually associated with better mathematics performance even if the new curriculum has no effect. The result is that causal inferences from QEDs are typically much weaker than those associated with RCTs unless the impact of selection bias is controlled. To produce the strongest possible causal inferences with QEDs, methodological standards recommend controlling for selection bias by adding predictors that treatment and control conditions may differ on such as SES in the data analysis, or employing matching procedures which are typically based on propensity scores (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007).

Regression discontinuity designs (RDDs) are increasingly used to assess the impact of treatment and control conditions when RCTs are not feasible. The basic RDD is a pretest-posttest two group (treatment, control) design. In a RDD, a variable which is typically a pretest is used to create a cutoff with cases with scores below and above the cutoff assigned to treatment and control groups (or vice versa). The groups are then compared on the change (discontinuity) in the pretest-posttest relationship at the cutoff (posttest is the outcome variable). The logic of RD is based on the crucial role of pretests in taking subject differences into account (Steiner, Cook, Shadish, & Clark, 2010), in that subjects with similar pretest scores can often be treated as approximately equal on background variables such as SES, enhancing causal inferences (Bloom, 2010).

For example, Jitendra, Harwell, Lm, Karl, and Slater (2018) used a RDD to examine the impact of an intervention designed to improve the mathematical problem-solving skills of students categorized as being at risk for having significant mathematical difficulties. A pretest measuring these skills (also referred to as the 'forcing' or 'running' variable) was used to categorize students as at risk for significant mathematical difficulties using a cutoff score of 9 which corresponded to the 35th percentile. Students with scores equal to or below 9 were at risk for significant mathematics difficulties and those above the cutoff had a modest risk of mathematical difficulties. In Jitendra et al. (2018) all students received the intervention because the research question was whether the intervention was more effective for students at risk of having significant mathematical difficulties, i.e., the independent variable was whether a student had a high or modest risk for significant mathematical difficulties; in other RDD applications an intervention would only be administered to students with pretest scores below (or above) the cutoff (e.g., Robinson, 2010).

Subjects can also be assigned to treatment and control conditions in RDDs based on collateral information. Consider a study assessing the impact of an intervention designed to reduce the number of student school suspensions for misbehavior. A school policy might be used to generate a cutoff, such as a policy in which students who are tardy five or more times are automatically suspended. Students who have been tardy five or more times are assigned to the treatment group and those tardy one to four times to the control group.

There are two types of RDDs: sharp RDD and fuzzy RDD. The former means the probability of being assigned to treatment changes from 0 to 1 (or 1 to 0) at the cutoff. A fuzzy RDD arises when the probability of being assigned to treatment changes from 0 to a value somewhat less than 1 (or a value somewhat greater than 0 to 1) at the cutoff. An example of a fuzzy RDD is when a researcher decides to assign subjects to treatment who were close to but slightly above the cutoff. Sharp RDDs are preferred because inferences about the treatment effect are clearer and analyses are simpler. The assumptions of RDDs needed to support causal inferences should be checked including an outcome variable that should be measured in the same manner for both treatment and control groups, student scores on the cutoff variable were not manipulated, and the relationship between the cutoff variable (e.g., pretest) and outcome variable (e.g., posttest) is linear; otherwise the analysis must include predictors capturing the nonlinearity (Smith, Levesque, Kaufman, & Strumpf, 2017, pp. 941-942).

Despite the absence of random assignment to treatment and control groups properly constructed RDDs support strong causal inferences. As noted above, it is often reasonable to assume subjects at the cutoff are quite similar in ways related to their performance on the posttest. Hence, in Jitendra et al. (2018) comparing the posttest scores of students scoring 9 on the pretest and students scoring 10 should produce an unbiased (or almost unbiased) estimate of the difference between those categorized as having a high risk of significant mathematics difficulties and those with a modest risk because students with similar pretest scores are likely to be similar on many other characteristics related to performance on the outcome variable. A small mean difference on the posttest implies the intervention is equally effective for students categorized as having a high versus a modest risk of significant mathematics difficulties, and a non-negligible posttest mean difference that the intervention had a differential effect on the two groups of students. Similarly, comparing the posttest scores of students with pretest scores of 8 or 9 against those of students with pretest scores of 10 or 11 should also produce unbiased or nearly unbiased estimates of the intervention effect, and so on.

It is important to mention correlational designs (CDs) which are common in educational research, for example, TIMSS (2015) and PISA (2017) data were obtained from a CD. CDs are non-experimental designs and are not part of the WWC (2017) standards because of their inability to support causal inferences. The defining characteristic of a CD is that a single group of subjects is measured on two or more variables whose relationship is examined, which is consistent with Ellis and Levy's (2009) and Creswell (2012) description of these designs as determining the presence and degree of the relationship between variables. The absence of treatment and control conditions means causal inferences for data obtained from such designs are extremely difficult to justify. The Council for Exceptional Children (2014) stated "identifying evidence-based practices involves making causal determinations, and causality cannot be reasonably inferred from correlational designs (p. 2)".

Examples of Common Methodological Challenges

Attrition is a common challenge that can cause or aggravate selection bias, which is the most severe methodological challenge in research design because it can undermine causal inferences. We provide a brief example of a methodological challenge for each research design and general advice for novice researchers.

In RCTs, to assess the impact of a year-long online method for teaching English to non-native English speakers in the seventh grade in the upper Midwest in the U.S., a researcher employs a RCT design in which each of N = 660 sampled students will be assigned at random to a treatment group in which students are given access to the online method of learning English, or a control group in which students are not given access to the online method (n = 330 per group). The outcome variable is a test of written English administered at the end of one-year. To assign students at random the SPSS (IBM Corp., 2011) computer program is used to generate N = 660 binary numbers (0, 1) such that 330 are 0s and 330 are 1s; students with a 1 are assigned to the treatment condition and those with a 0 to the control condition. An important challenge in RCTs is ensuring that whatever attrition exists is similar in treatment and control groups (WWC, 2017). If 5% of treatment students and 5% of control students are lost (missing data) then estimates of an intervention effect can often still be treated as unbiased (or almost unbiased); if 20% of treatment students and 5% of control students are lost estimates of an intervention effect are more likely to be biased.

In QEDs, using the same information provided in the previous example, a researcher turns to schools already applying the new online method for teaching English to non-native speakers of English and defines n = 330 students in these schools as representing the treatment group. Another 330 students were obtained from schools not using the online method and represent the control group. The most severe methodological challenge is selection bias meaning the treatment and control groups differ on the outcome at the study onset due to variables that confound (bias) comparisons. For example, the treatment group may have a higher (or lower) percentage of students in poverty, more (or less) access to high quality internet service, or stronger (or weaker) English teaching all of which affect the outcome. Collecting data on potentially confounding variables that can bias treatment and control comparisons is critical for controlling their effects (e.g., by including these variables as predictors in data analyses or employing matching procedures based on propensity scores).

In RDDs, using the information provided above, the researcher decides to utilize a RDD in the study of an online method to teach English to non-native English speakers. A test of written English administered at the beginning of the school year to non-native English speaking seventh grade students serves as the forcing variable, i.e., students scoring below a specified value, such as the score corresponding to the 35th percentile, are assigned to the treatment group (given access to the online method for learning English) or are assigned to the control group (no access to the online method for learning English). The test of written English administered at the end of one year serves as the posttest.

One important methodological challenge in RDDs is attrition which should be tracked in the treatment and control groups. Another challenge is choosing a forcing variable and associated cutoff that provides results that speak to the research question motivating the study. For example, Jitendra et al. (2018) chose a pretest as the forcing variable and a score corresponding to the 35th percentile as the cutoff based on considerable prior research results for identifying students at risk of significant mathematical difficulties. A different cutoff choice by these authors, for example, a pretest score corresponding to the 50th percentile, would still be expected to produce an unbiased (or almost unbiased) estimate of a treatment effect but the choice of a non-literature-based cutoff would likely undermine the usefulness of study results.

In CDs, the researcher is interested in investigating the relationship between the written English proficiency of seventh grade students who are non-native English speakers enrolled in an online method for learning English, and variables such as students' grade point average and score on a standardized test of English literacy. The researcher continues to sample students who have been using the online method until N = 660 are obtained, who are subsequently administered the test of written English. Several methodological challenges are linked to this design but the most important is not having treatment and control groups, which means the effectiveness of the online method cannot be directly studied.

Practical Advice

We offer five pieces of advice on research design to the novice researcher:

1. Producing unbiased estimates of effects should be the goal of quantitative studies. The recommended standard is randomly assigning subjects to treatment and control conditions because this should produce unbiased estimates of a treatment effect and support strong causal inferences.

2. The use of an RCT does not guarantee unbiased estimates and strong causal inferences unless factors that can undermine such inferences are controlled. The WWC (2017) stated that researchers should be particularly concerned about attrition (missing data). It is a good idea to report the percentage of missing data in the treatment and control groups as well as for each variable. Differential attrition refers to a difference in the attrition rates for the treatment and control groups and can represent a severe threat to valid inferences. Every effort should be made to minimize attrition. Remedies such as data imputation, which produce complete data, require rigorous assumptions be met.

3. In many settings, random assignment is impossible and QEDs comparing existing groups are used. Such designs are prone to selection bias in which groups a priori

differ on an outcome due to the presence of confounding variables that are not equally distributed across the groups and are correlated with an outcome. Every effort must be made to eliminate this bias, which typically takes the form of statistical control through the use of predictor variables in regression models or matching using propensity scores (Schneider et al., 2007).

4. RDDs represent a powerful tool that can produce unbiased (or nearly unbiased) estimates and support strong causal inferences. A key to these designs is employing a forcing variable (e.g., pretest) that through an appropriately chosen cutoff score creates treatment and control groups. In general, a RDD should be employed before a QED if at all possible.

5. Methodological standards provide no place for CDs because they cannot support strong causal inferences. These designs should generally be limited to preliminary research studies in which the goal is to provide empirical evidence of relationships among key variables.

Instrumentation

The importance of using accurate instruments (tests, surveys, questionnaires) with exemplary psychometric properties in the educational sciences is well documented (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Biemer & Lyberg, 2003; Danner et al., 2016; Fry, 1960; Kane, 2001, 2013; White, Carey, & Dailey, 2001). We employ the definition of instrumentation provided by Hsu and Sandford (2010):

"Instrumentation refers to the tools or means by which investigators attempt to measure variables or items of interest in the data-collection process. It is related not only to instrument design, selection, construction, and assessment, but also to the conditions under which the designated instruments are administered—the instrument is the device used by investigators for collecting data." (p. 608).

Measurement instruments represent the foundation of empirical research in the educational sciences (Danner et al., 2016) yet selecting, modifying, or constructing instruments that support reliable and valid interpretations of scores is challenging, likely more so for novice researchers. The goal of this paper is to encourage novice researchers to employ available standards for instruments (e.g., AERA, APA, NCME, 2014; Council for Exceptional Children [CEC], 2014; WWC, 2017).

Psychometric Properties of Measurement Instruments

The two critical properties of measurement instruments are the reliability and validity of interpretations and uses and consequently inferences made using instrument scores

(Haladyna, 2004; Haladyna & Rodriguez, 2013; Kane, 2013, 2016; Linn, 2006). The need for instruments with strong psychometric properties, and the difficulty of constructing such instruments, is well known (WWC, 2017). Unsurprisingly, educational research is plagued by poor instruments, in large part because of a belief that constructing psychometrically strong instruments is a modest task when the opposite is true.

Reliability and validity standards are typically applied to outcome variables in a study but are often relevant for other variables such as predictors in regression analyses. For example, in Harwell et al. (2009) student socio-economic status (SES), high school grade point average in mathematics classes, and year enrolled served as predictors. Applying instrumentation standards to these variables speaks to the reliability and validity of inferences based on these variables.

Reliability is not an absolute property of an instrument but rather refers to the amount of measurement precision (consistency) (Reynolds & Livingston, 2012) of scores. Reliability coefficients range between 0 to 1 or -1 to 1 depending on the coefficient, and instruments with more items should generally produce higher reliability because they provide more information. For example, suppose the reliability coefficient of a test intended to measure the construct proficiency with fractions equals .90. This value means that 90% of the variation in test scores is due to the construct and 10% reflects measurement error.

Another interpretation of reliability is the following: If a sample of subjects are given the same instrument twice and the rank-order of scores is quite similar across assessments, the reliability of inferences based on scores is high. In this case, the subject with the highest score at time 1 also likely obtained the highest score at time 2 (it is not necessary that subject 1 obtain the same score on both assessments), the subject with the second-highest score at time 1 also is likely to have the second-highest score at time 2 (it is not necessary that subject 2 obtain the same score on both assessments), and so on. On the other hand, if a sample of subjects is given the same instrument twice and the rank-order of scores is quite different across assessments reliability will be low, for example, the subject with the highest score at time 1 likely does not obtain the highest score at time 2, the subject with the second-highest score at time 1 likely does not obtain the second-highest score at time 2, and so on.

Several measures of reliability are available. If an instrument is administered twice a Pearson product-moment correlation typically serves as a reliability measure. If a single assessment is used, Cronbach's alpha is popular and is available in most computer programs performing data analysis. However, Cronbach's alpha has significant deficiencies (Dunn, Baguley, & Brunsden, 2014; Green & Hershberger, 2000; Raykov, 2001; Tang & Cui, 2012; Yang & Green, 2011; Zhang & Yuan, 2016; Zimmerman, Zumbo,

& Lalonde, 1993; Zinbarg, Revelle, Yovel, & Li, 2005). Even Cronbach expressed concerns over the usefulness of alpha as it covers only a modest percentage of measurement uses for which reliability information is needed (Cronbach & Shavelson, 2004). A key assumption of coefficient alpha is that all items on an instrument measure the same construct (e.g., mathematics achievement) and possess the same factor analysis loadings which capture the relationship between each item and the construct (the essentially tau-equivalent model), which may not occur in practice An arguably more realistic measure of reliability is the omega coefficient which also assumes items measure the same construct but allows factor loadings to vary (The congeneric model) (McDonald, 1999). Omega is available in R (R Core Team, 2018) and the jMetrik software (Version 2.1.0; Meyer, 2011) both of which can be downloaded without charge.

Reliability coefficients $\geq$ .80 are often acceptable (CEC, 2014) although a variety of factors impact minimally acceptable values, such as the characteristic or trait being measured and the purpose of the instrument. For example, a reliability of .70 would likely be unacceptable for a test of mathematics achievement used in making college admission decisions but perfectly acceptable for a questionnaire being developed to assess attitudes towards public education. An infrequently used but promising approach to determine minimally acceptable reliability is to employ a decision criterion (Gugiu & Gugiu, 2018).

Validity is the accuracy with which inferences about a subject's status are made based on their score. Like reliability, validity is not an absolute property of an instrument but refers to the proposed interpretations and uses of scores. Chan (2014) summarized this perspective: "Validity is about the inferences, claims, or decisions that we make based on instrument scores, not the instrument itself" (p. 10). The AERA, APA, and NCME (2014) instrumentation standards provided a similar definition:

"Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself." (p. 11)

Validity is best thought of as the single overall judgment of the adequacy and accuracy of an instrument's interpretation or intended use. Several types of validity evidence are available but four are especially prominent and often overlap: face validity, content validity, criterion validity, and construct validity. The WWC (2017) standards emphasized the importance of face validity evidence, which reflects the extent to which an instrument appears to do what it claims. For example, a test of written English proficiency that required U.S. students to interpret political cartoons from the United Kingdom in order to complete items would likely be viewed as lacking face validity.

Content validity evidence refers to the extent to which an instrument reflects relevant facets of an underlying construct like proficiency with fractions. Evidence of content validity is often provided by a logical evaluation of the degree to which items cover relevant facets (often called an instrument's blueprint), for example, the extent to which items reflect the steps needed to solve fraction problems. Both face validity and content validity evidence typically includes the judgments of experts who represent the domain of the intended uses of the instrument such as teachers, college admissions officers, or public policy staff.

Criterion validity evidence can be concurrent or predictive. Concurrent validity evidence assesses the relationship between an instrument and an existing measure which ideally has strong psychometric properties. Predictive validity evidence reflects how well scores predict a future outcome, for example, the extent to which scores of seventh grade students on a test of fractions predict their performance on an algebra test in eighth grade. Criterion validity evidence can be assessed by estimating the correlation coefficient between an instrument and current (concurrent) and later (predictive) performances.

Construct validity evidence reflects the extent to which inferences about a construct such as proficiency with fractions is accurate. Evidence of construct validity is typically both theoretical and empirical. The former reflects the expected structure of instrument items (e.g., all items reflect a single construct depicted in a theoretical model and test blueprint), and the latter takes the form of non-negligible correlations between instrument scores and variables these scores are expected to be related to (e.g., student scores on a test of fractions and teacher ratings of students' mathematics proficiency). Construct validity evidence can also be generated using factor-analytic techniques. Deciding which type(s) of validity evidence to obtain and report is a critical decision and should be based on the intended interpretations and uses of instrument scores.

The relationship between reliability and validity is important but somewhat confusing because they are distinct yet related concepts. An instrument supporting reliable inferences may not support valid inferences. For example, a test of proficiency with fractions may produce consistent scores over repeated test administrations but not cover the domain it purports to or fail to predict future performance on an algebra test. However, a test with validity evidence must possess reliability because valid inferences about scores must be consistent.

In sum, measurement instruments with strong reliability and validity evidence strengthen inferences based on instrument scores, whereas those with weak reliability and validity evidence can undermine inferences (Hsu & Sandford, 2010). The following sections illustrate the options available for selecting, modifying, or creating instruments, recommendations for instrument administration and data collection, common challenges using instruments and how to respond to them, and practical advice targeting novice researchers.

Options for Instruments Used in Data Collection

Researchers have three options for instruments: Select a published (existing) instrument in its current form, modify an existing instrument, or construct a new instrument (Creswell, 2012). This choice is often a critical methodological decision that can significantly enhance or undermine the quality of a study's findings. We urge researchers to employ the above order in practice, i.e., select an existing instrument with evidence of strong psychometric properties if possible, modify an existing instrument with evidence of strong psychometric properties if necessary, and construct an instrument as a last resort.

Gay and Airasian (2000) outlined seven factors to consider in selecting an existing instrument: "(1) The name, publisher, and cost, (2) a brief description of the purpose of the instrument, (3) validity and reliability data, (4) the group for whom the instrument is intended, (5) administration requirements, (6) information regarding scoring and interpretation, and (7) reviewers' overall impressions" (p. 145). We encourage novice researchers to begin their search for an existing instrument with the "*Mental Measurement Yearbook"* (MMY) published by The Buros Center for Testing at the University of Nebraska-Lincoln. These yearbooks appear every 3-5 years and contain reviews by professional educators of hundreds of instruments in the educational sciences (Carlson, Geisinger, & Jonson, 2017). The reviews contain information about the purpose, population, publication dates, administration time, score scale, price, technical issues, and psychometric properties of instruments. The instruments reviewed assess a wide range of content, for example, mathematics, reading, and science achievement, auditory perceptual skills, behavior assessment, foreign language proficiency, job related skills, intelligence and aptitude, personality traits, and teaching quality. Instruments with positive MMY reviews that meet the needs of a study deserve careful consideration whereas those with negative reviews should generally be avoided.

A second option is modifying an existing instrument. Modifications often take the form of simplifying the wording of directions and items, adding or deleting items, and extending or shortening the time to complete the instrument. Modifications should be made after obtaining the approval of the instrument author(s) and publisher which can be a lengthy and complex process. In addition, the psychometric properties of the original instrument do not necessarily apply to the modified instrument.

A third option is constructing a new instrument. Guidance in developing and validating an instrument is available through the standards published jointly by AERA, APA, and NCME (2014). A careful review of the steps of constructing an instrument detailed in these standards shows that developing and validating a psychometrically strong instrument is a complex and time-consuming task that can take years. If a researcher designs their own instrument the rationale for doing so should be provided and should include evidence (a) of a paucity of existing instruments available for measuring the construct(s) of interest in a study, (b) that the new instrument represents a significant contribution to a field, (c) that the design and validation of the instrument is consistent with recommended standards.

Below is a summary of the steps for constructing a new measurement instrument laid out in the AERA, APA, and NCME (2014) standards:

- Determine the purpose and rationale for designing a new instrument and the intended interpretations and uses of scores as well as the target population.
- Review related literature to provide deep insight into the construct(s) being measured.
- Operationally define the construct(s) and their sub-dimensions (components) if relevant (i.e., construct a test blueprint).
- Construct an item pool that covers relevant facets of a construct. An important decision in this process is item format, for example, multiple choice, open-ended items, or rating scales.
- Ask content experts to review and revise the item pool as needed given the instrument's purpose, intended interpretations, and target population.
- Select items from the item pool to comprise the first draft of the instrument based on their content coverage, readability, and fairness.
- Utilize the "thinking aloud" strategy among those constructing the instrument and a small group of examine in which reactions to the instrument are shared to further refine the first draft.
- Plan to administer the initial draft in a pilot study in a way that ensures data collection bias will be minimized (date, time, sampling, scoring, data collector characteristics).
- Score responses from the pilot study using theoretically-grounded rubrics to score open-ended items if present.

- Estimate the psychometric properties (reliability and validity).
- Refine the instrument based on the results of the pilot study.
- Repeat the above steps as needed. In many instances, an instrument is not ready for use after a single piloting.

In sum, researchers ideally select an existing instrument that possesses evidence of strong psychometric properties or modify an existing instrument. Researchers who choose to construct an instrument should follow the AERA, APA, and NCME (2014) guidelines for doing so and should recognize that the process of developing and piloting an instrument until it shows strong reliability and validity is likely to take considerable time, effort, and resources.

Recommendations for Instrument Administration and Data Collection

Once an instrument is available for data collection in a study, it is important to minimize factors that can undermine data quality. McMillan and Gogia (2014) pointed out the value of researchers carefully selecting the conditions, time, and place for administration to help ensure data quality. Factors like respondent fatigue and mood, environment noise, or test monitors who provide confusing instructions for completing the instrument can produce inaccurate responses, missing data, and random measurement errors which lower reliability. Additional problems that can undermine data quality may emerge if the instrument appears on a computer or tablet, such as internet connection or firewall issues, equipment shortcomings (e.g., unresponsive mouse), or difficulties in the computer program used to deliver the instrument and record responses (e.g., the computer program "freezes"). If an instrument requires raters or observers to assess respondents' performance, it is important that raters be trained such that variability among their ratings is minimized to enhance reliability and validity of inferences (Harwell, 1999).

Examples of Common Challenges in Instrumentation

Various challenges may arise in instrumentation. Perhaps one of the more common challenges is whether an instrument positively reviewed in the MMY should be used in a study with a somewhat different purpose or target population than that of the positively-reviewed instrument. For example, a researcher may use the MMY to identify an existing test that assesses the ability of sixth grade students to solve fraction problems, but discovers that only a test of fractions for fourth and fifth grade students (FRACT) is positively reviewed. Should the researcher administer FRACT to sixth grade students, modify FRACT to make it appropriate for sixth grade students, or develop a new test of fractions that targets sixth grade students? Other instrumentation challenges include employing a sample consistent with the target population and the availability of individuals with psychometric expertise to analyze item response data and interpret the results. Below we expand on these challenges in the context of selecting, modifying, or developing a new instrument. FRACT is the example used to illustrate such challenges.

Suppose a researcher turned to the MMY for a published test measuring the proficiency of sixth grade students' proficiency with fractions. The MMY produced a positively-reviewed test (FRACT) designed for fourth and fifth grade students but none for sixth grade students. FRACT was published in 2010 after a lengthy development process that followed AERA, APA, and NCME (2014) recommendations, consist of multiple choice and open-ended items, and has evidence of strong reliability and construct validity. If the researcher administers FRACT to sixth grade students, the reliability and validity of test-based inferences may be incorrect, but if the researcher decides to not use FRACT in its published form either FRACT needs to be modified to accommodate sixth grade students or a new instrument needs to be developed.

Continuing with the above example, suppose the researcher decides to modify FRACT to ensure it is appropriate for sixth grade students. The modifications include revising the wording in test directions and test items, increasing the difficulty of items, and reducing the time allotted to complete the test. The researcher contacts the FRACT author(s) and test publisher for permission to modify the existing instrument. After months of negotiations over the modifications, the author(s) and publisher agree to the modifications, but are concerned that the strong psychometric properties of the original test may not transfer to the modified version. To respond to this concern, the researcher plans to follow the recommended steps of the AERA, APA, and NCME (2014) test standards, and to provide evidence the properties of the modified and original instruments are similar by piloting the modified instrument and reporting reliability and validity evidence.

Suppose instead the researcher plans to construct a new instrument (FRACT2) that targets the proficiency with fractions of sixth grade students. The researcher begins by examining the recommendations in the AERA, APA, and NCME (2014) standards. If the researcher does this on their own it is likely to take several months, a year, or longer; if a team of researchers are involved the process presumably takes less time. The steps are:

- Clarify the purpose of the test and the target population (e.g., measuring the proficiency with fractions of sixth grade students).

- Review literature related to assessing the proficiency with fractions of sixth grade students to identify the desired skills that should be covered in the test (test blueprint) such as the ability to recognize common underlying problem structures, represent problems using visual-schematic diagrams, plan how to solve problems, and solve and check the reasonableness of answers.

- Construct test items that will elicit the desired skills. The researcher decides that 20 multiple choice items and 10 open-ended tasks will appear on the test. Typically more items are developed than appear on the final version of the instrument because piloting the instrument will likely reveal items that perform poorly and should be discarded. As a result 30 multiple choice items and 15 open-ended items are developed. It is important to construct items that avoid culturally dependent or insensitive language (e.g., a fraction problem involving the number of rooms in a student's household when some students may be homeless or what constitutes a room in a household may vary by culture).

- Identify content experts (e.g., teachers, faculty researchers) who review and revise the items in terms of readability and suitability for sixth grade students. This feedback prompts revision of the items and the development of a scoring rubric for open-ended items.

- Utilize a "thinking aloud" strategy in which a group of 15 sixth grade students read the test directions and items and provide feedback on their readability and what they believe is an appropriate response. This feedback is used to further revise the test.

- Pilot an initial version of the test using a sample of N = 300 sixth grade students. This sample size should be large enough to ensure pilot data can be confidently used to estimate the psychometric properties of the test along with information about the adequacy of the testing protocol (e.g., directions given to students, time allotted for completing the test).

- Score student responses from the pilot study and compute reliability using the omega coefficient which allows items to have different loadings on the factor "proficiency with fractions" and supports different item formats (binary scoring, open-ended scoring) in the same instrument. Correlate test scores with collateral information such as scores from a standardized test of mathematics to provides evidence of criterion validity.

- Develop a transparent set of guidelines for keeping or omitting items based on the results of the pilot study. For example, the pilot data may suggest 12 multiple-choice items and five open-ended items should be discarded. At this point, one of two things happens: (a) The researcher decides the instrument has satisfactory psychometric properties (e.g., omega coefficient of the remaining 18 + 10 = 28 items is .86, correlation of standardized mathematics scores and FRACT2 scores is .45) and should be used to collect data in the main study, or (b) The researcher decides the instrument has unsatisfactory psychometric properties (e.g., omega coefficient of the 28 remaining items is .62 and additional items need to be written, correlation of standardized mathematics scores and FRACT2 scores is .12). In this case, the test must go through another round of development and piloting.

Practical Advice

We offer three pieces of advice on instrumentation challenges to novice researchers:

1. The impact of instruments on study-based inferences as a function of their reliability and validity evidence speaks to the need to employ instruments with strong psychometric properties. Put bluntly, it is difficult to over-emphasize the damage a poor instrument can do to inferences from otherwise well-designed studies. The use of existing instruments with compelling evidence of strong psychometric properties is always preferred assuming the instrument is consistent with a study’s purpose, intended interpretations, and target population. On the other hand, a well-reviewed existing instrument that does not meet these conditions should not be used.

2. Modifying an existing instrument is typically preferable to constructing one but the possibility the modified test performs differently from the original in ways that affect its psychometric properties should be acknowledged and studied. This will likely involve following several steps of the AERA, APA, and NCME (2014) standards such as piloting the modified instrument.

3. Researchers should not construct an instrument for a study unless it is absolutely necessary to do so, and in this case the AERA, APA, and NCME (2014) standards for test construction should be followed to help ensure the instrument possesses strong psychometric properties. This is often a time consuming process that even under ideal circumstances will require significant resources and take several months to more than one year.

The Extent to Which a Sample of U.S. and Non-U.S. Quantitative Education Studies Capitalize on Recommended methodological Standards

The paper’s premise that researchers often do not capitalize on recommended sampling, research design, instrumentation and standards was assessed using a sample of U.S. and non-U.S. published studies in education. For the U.S. studies, we sampled studies appearing in the American Educational Research Journal (AERJ), the flagship journal of AERA, in 2016 (Vol. 53) and 2017 (Vol. 54). For non-U.S. studies we surveyed the International Journal of Educational Research (IJER) published by the European

Association for Research on Learning and Instruction in 2016 (Vol. 75 - Vol. 80) and 2017 (Vol. 81 - Vo. 86). IJER is an international outlet for educational research conducted in Africa, Asia, Australia, and Europe.

For sampling, a study was coded 4 if all participants were selected at random, 3 if some participants were selected at random (e.g., random sampling of schools but not students within schools), 2 if the study employed a non-probability-based sampling such as a convenience sample, and 1 if the sampling mechanism was not clearly stated. A similar strategy was followed to assess the research design utilized in a study: a coded value of 4 means all subjects were randomly assigned to groups (e.g., treatment and control), 3 was assigned if the study employed a quasi-experimental design (non-random assignment of subjects to groups), 2 if the study employed a correlational research design (one group of subjects), and 1 if the research design was not clearly stated. Concerning instrumentation, a study was coded 4 if an instrument was international, national, or regional in its origin and purpose because these instruments typically report following test construction standards and demonstrate stronger psychometric evidence. For example, a study using an instrument constructed for an international assessment such as PISA or TIMSS, a national test such as the American College Testing (ACT) assessment, or a state-mandated test was coded 4. A value of 3 was assigned if study-based evidence of strong psychometric properties of an existing instrument or a new instrument developed for a study was reported, 2 if study-based evidence of the psychometric properties of an existing instrument or a new instrument developed for a study were weak, and 1 if psychometric evidence for an instrument was not clearly reported.

A total of 111 studies were reviewed (52 from the AERJ, 59 from IJER). One of the authors did the coding but both authors discussed coding issues in particular studies (e.g., should a study be coded 1 or 2 for research design given the limited information provided) until a consensus was reached. The results in Table 1 indicate that the methodological standards of sampling and research design received minimal attention whereas instrumentation received moderate attention. For studies appearing in *AERJ*, overall 30.8% utilized national or international datasets that employed random sampling and 3.8% employed randomization at the cluster level but non-random sampling at level 1. Non-random sampling was used in 65.4% of the studies reviewed in AERJ, suggesting that generalizing results for these studies is challenging. For research design in the AERJ studies, 7.7% employed RCTs, 25% employed QEDs, and 67.3% employed CDs. Overall 55.8% of AERJ studies used international, national, or state-mandated instruments, approximately one-third of these studies (32.7%) used existing instruments or created instruments showing strong psychometric properties, and 11.5% used newly created instruments but did not report evidence of their psychometric properties.

On the whole, studies appearing in *IJER* showed less evidence of capitalizing on recommended methodological standards. Overall 10.7% of the studies utilized random sampling (most as a result of using PISA or TIMSS data) and 79.7% used non-random sampling. For research design 11.9% employed RCTs, 25.4% employed QEDs, and 62.7% employed CDs. Overall 30.5% of these studies used international or national instruments with evidence of strong psychometric properties, 37.3% used existing instruments or created an instrument for which evidence of strong psychometric properties was reported. Approximately one-third of the studies (32.2%) used existing instruments or created new instruments for which the reported psychometric evidence was weak or not reported at all.

Table 1. Frequency of Articles Employing the Three Methodological Components by Journal and Year (N= 111)

Methodo-logical Components	Journal	American Educational Research Journal						International Journal of Educational Research					
	Year	2016 N= 32		2017 N= 20		Total N= 52		2016 N= 27		2017 N= 32		Total N= 59	
	Code label (Value)	N	%	N	%	N	%	N	%	N	%	N	%
Sampling	Fully random (4)	13	40.63	3	15	16	30.77	2	7.41	4	12.5	6	10.17
	Partially random (3)	2	6.25	---	---	2	3.85	3	11.11	3	9.38	6	10.17
	Non-random (2)	17	53.12	17	85	34	65.38	22	81.48	25	78.12	47	79.66
	Not stated/ unclear (1)	---	---	---	---	---	---	---	---	---	---	---	---
Research Design	Randomized (4)	2	6.25	2	10	4	7.69	2	7.41	5	15.63	7	11.86
	Quasi-experimental (3)	5	15.63	8	40	13	25	9	33.33	6	18.75	15	25.42
	Correlational (2)	25	78.12	10	50	35	67.31	16	59.26	21	65.62	37	62.72
	Unstated/ unclear (1)	---	---	---	---	---	---	---	---	---	---	---	---
Instrumentation	Nat., Int., or SM: SPER (4)	21	65.62	8	40	29	55.77	6	22.22	12	37.5	18	30.51
	Existed or created: SPER (3)	5	15.63	12	60	17	32.69	12	44.45	10	31.25	22	37.29
	Existed or created: WPER (2)	6	18.75	---	---	6	11.54	9	33.33	10	31.25	19	32.20
	Existed or created: Psychometric evidence was not clearly reported (1)	---	---	---	---	---	---	---	---	---	---	---	---

Note. Int.= International, Nat.= National, SM= State-mandated, SPER= Strong psychometric evidence reported

Conclusion

Strengthening inferences of quantitative education studies is a critical goal. The current paper speaks to this goal by providing novice researchers with descriptions of three facets of quantitative methodology (sampling, research design, instrumentation) that may not receive the attention they deserve, along with examples and practical advice that should promote stronger inferences. A review of a sample of U.S. and non U.S. studies provided evidence methodological standards for sampling and research design are under-capitalized. Accordingly, this work should be of value to novice researchers planning to conduct quantitative studies. Audiences for this paper include new faculty, individuals beginning non-faculty roles such as a working in a university-affiliated research center or a government-funded education center, faculty transitioning to more research-oriented work, and graduate students conducting their own research. This paper could also be used for instructional purposes in research methodology classes.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*(9), 839-851. https://doi.org/10.1037/0003-066X.63.9.83

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force report. *American Psychologist, 73*, 3–25. http://dx.doi.org/10.1037/amp0000191

Battalgia, M. P. (2008). Nonprobability sampling. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (PP. 1-7). Thousand Oaks, CA: Sage Publications, Inc.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken, NJ: John Wiley & Sons.

Bloom, H. (2010). *Modern regression discontinuity analysis*. New York, NY: MDRC.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally & Company.

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.) (2017). *The twentieth mental measurement yearbook*. Lincoln, NE: The University of Nebraska Press.

Chan, E. K. H. (2014). Standards and guidelines for validation practices: Development and validation of measurement instruments. In B. D. Zumbo, & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 9-24). Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9_2

Council for Exceptional Children (2014). Standards for evidence-based practices in special education. Arlington, VA: Council for Exceptional Children.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418. https://doi: 10.1177/ 0013164404266386

Danner, D., Blasius, J., Breyer, B., Eifler, S., Menold, N., Paulhus, D. L., . . . Ziegler, M. (2016). Current challenges, new developments, and future directions in scale construction. *European Journal of Psychological Assessment, 32*(3): 175–180. http://dx.doi.org/10.1027/1015-5759/a000375

Dedrick, R. F., Ferron, J. M., Hess, M. R. Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*(1), 69-102. https://doi.org/10.3102/0034654308325581

Dunn, T. J., Baguley, T. & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399-412. https://doi: 10.1111/bjop.12046

Ellis, T. J., & Levy, Y. (2009). Towards a guide for novice researchers on research methodology: Review and proposed methods. *Issues in Informing Science and Information Technology, 6*, 323-337.

Ellis, T. J., & Levy, Y. (2010). *A guide for novice researchers: Design and development research methods*. Proceedings of Informing Science & IT Education Conference.

Fath, K. Q. (2014). Reporting methods and analyses in higher education research: Hierarchical linear and OLS regression models. Unpublished dissertation, Loyola University, Chicago, IL.

Fry, E. B. (1960). Research tools: Instrumentation in educational research. *Review of Educational Research, 30*(5), 513-521.

Gay, L. R., & Airasian, P. (2000). *Educational research: Competencies for analysis and application* (6th ed.). Upper Saddle River, NJ: Printice-Hall, Inc.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(2), 251-270. https://doi: 10.1207/ S15328007SEM0702_6

Gugiu, C., & Gugiu, M. (2018). Determining the minimum reliability standard based on a decision criterion. *The Journal of Experimental Education, 86*(3), 458-472. https://doi.org/10.1080/00220973.2017.1315712

Haladyna. T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Harwell, M. R. (1999). Evaluating the validity of educational rating data. *Educational and Psychological Measurement*, *59* (1), 25-27. https://doi org/10.1177/0013164499591002

Harwell, M. R. (2011). Research design: Qualitative, quantitative, and mixed methods. In C. F. Conrad, & R. C. Serlin (Eds.), *The Sage handbook for research in education: Pursuing ideas as the keystone of exemplary inquiry* (2nd ed.) (pp. 147-164). Thousand Oaks, CA: Sage Publication Inc.

Harwell, M. R., Post, R. T., Cutler, A., Maeda, Y., Anderson, E., Norman, K. W., & Medhanie, A. (2009). The preparation of students from national science foundation–funded and commercially developed high school mathematics curricula for their first university mathematics course. *American Educational Research Journal, 46*(1), 203-231. https://doi.org/10.3102/0002831208323368

Hsu, C., & Sandford, B. A. (2010). Instrumentation. In N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 608-610). Thousand Oaks, CA: Sage Publication Inc.

IBM Corp. (2011). *IBM SPSS Statistics for Windows* (Version 20.0) [Computer software]. Armonk, NY: IBM Corp.

Jitendra, A. K., Harwell, M. R., Lm, S., Karl, S. R., & Slater, S. C. (2018). Using regression discontinuity to estimate the effects of a tier 1 research-based mathematics program in seventh-grade. *Exceptional Children, 85*(1), 46-65. https://doi.org/10.1177/0014402918784541

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational*

Measurement, 38(4)¸ 319-342. https:// dx.doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. http://dx.doi.org/10.1111/jedm.12000

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198-211. https://doi.org/10.1080/0969594X.2015.1060192

Kleinman, K. (2017). Cluster-randomized trials. In C. Gatsonis, & S. C. Morto (Eds.), *Methods in comparative effectiveness research* (pp. 131–155). Boca Raton, FL: Taylor & Francis Group, LLC.

Linn, R. L. (2006). The standards for educational and psychological testing: Guidance in test development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp.27-38). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Boston, MA: Cengage Learning.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

McMillan, J. H., & Gogia, L. (2014). Data collection in educational research. Oxford Bibliographies in Education. doi: 10.1093/obo/9780199756810-0087

Meyer, J. P. (2011). *jMetrik: Open source psychometric software* [computer program]. Retrieved from www.ItemAnalysis.com

Pedhazur, E. J., & Schmelkin, L.P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Raykov, T. (2001). Bias of coefficient alpha for congeneric measures with correlated errors. *Applied Psychological Measurement, 25*(1), 69-76. http://dx.doi: 10.1177/01466216010251005

R Core Team (2018). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org

Reynolds, C. R., & Livingston, R. B. (2012). *Mastering modern psychological testing: Theory & methods*. Boston, MA: Pearson.

Robinson, J. P. (2010). The effects of test translation on young English learners' mathematics performance. *Educational Researcher, 39*(8), 582–590. https://doi.org/10.3102/0013189X10389811

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs*. Washington, DC: American Educational Research Association.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inferences*. New York, NY: Houghton Mifflin Company.

Smith, L. M., Levesque, L. E., Kaufman, J. S., & Strumpf, E. C. (2017). Strategies for evaluating the assumptions of the regression discontinuity design: A case study using a human papillomavirus vaccination programme. *International Journal of Epidemiology, 46*(3), 939–949. https://doi.org/10.1093/ije/dyw195

Stata Corp. (2015). *Stata Statistical Software: Release 14* [Computer software]. College Station, TX: StataCorp LP.

Steiner, P., Cook, T. D, Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*, 250–267. doi:10.1037/a0018719

Tang, W., & Cui, Y. (2012, April). *A simulation study for comparing three lower bounds to reliability.* Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.

Trochim, W. M. K., & Land, D. A. (1982). Designing designs for research. *The Researcher, 1,* 1–6.

U.S. Department of Education (2013). Common guidelines for education research and development. A Report from the Institute of Education Sciences. Department of Education's Institute of Education Sciences, Washington, D.C.

White, J. A., Carey, L. M., & Dailey, K. A. (2001). Web-based instrumentation in educational survey research. *WebNet Journal*, 46-50.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604. https://doi.org/10.1037/0003-066X.54.8.594

WWC standards (2017). *What works clearinghouse procedures and standards handbook* (Version 4). U.S. Department of Education's Institute of Education Sciences.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, 29, 377–392. http://dx.doi: 10.1177/ 0734282911406668

Zhang, Z., & Yuan, K. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement, 76*(3), 387–411. http://dx.doi: 10.1177/0013164415594658

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53*(1), 33-49. http://dx.doi: 10.1177/0013164493053001003

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωH: There relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123-133. http://dx. doi: 10.1007/s11336-