Education Leaders' Knowledge of Causal Research Design: A Measurement Challenge

Heather C. Hill, Harvard Graduate School of Education

Derek C. Briggs, University of Colorado

Author's note: This study was funded by the Institute of Education Sciences (10.13039/100005246, R305C140008).

Abstract

Federal policy has both incentivized and supported better use of research evidence by educational leaders. However, the extent to which these leaders are well-positioned to understand foundational principles from research design and statistics, including those that underlie the What Works Clearinghouse ratings of research studies, remains an open question. To investigate educational leaders' knowledge of these topics, we developed a construct map and items representing key concepts, then conducted surveys containing those items with a small pilot sample (n=178) and a larger nationally representative sample (n=733) of educational leaders. We found that leaders' knowledge was surprisingly inconsistent across topics. We also found most items were answered correctly by less than half of respondents, with cognitive interviews suggesting that some of those correct answers derived from guessing or test-taking techniques. Our findings identify a roadblock to policymakers' contention that educational leaders should use research in decision-making.

Introduction

For more than a decade, federal policy has both incentivized and supported better use of research evidence by those responsible for establishing and enacting educational policies and programs. No Child Left Behind, the major Bush-era funding stream for schools, required local educational agencies (LEAs) to spend their federal dollars on programs backed by "scientifically-based" research. The NCLB reauthorization, completed in 2015 and titled Every Student Succeeds Act (ESSA), added detail, specifying tiers of evidence (e.g., promising to strong) and the types of studies (e.g., correlational, experimental) required for an intervention to reach each tier. The federal government has also invested in making research results more accessible, for instance through the Institute for Educational Sciences practice guide series and the *What Works Clearinghouse* (*WWC*) website (https://ies.ed.gov/ncee/wwc/PracticeGuides and https://ies.ed.gov/ncee/wwc/, respectively). In fact, educational practitioners, whom we define for this article as those in leadership roles at either a district or school level, indicate that they do use research to make decisions and expand their understanding of issues. For example, among those involved in adopting, eliminating, and designing programs, roughly four in five reported frequently using evidence as part of these activities (Authors et al., 2017).

However, the extent to which these practitioners are well-positioned to understand foundational research design principles, including those that underlie the ratings in WWC evaluations of research studies, remains an open question. Reading original research studies requires knowledge of the inferences and generalizations that are plausible on the basis of specific research designs. As well, research studies often use technically complex language and metrics, referencing quasi-experiments, power analyses, statistical significance and standard deviation units without explaining the meaning of these terms or why they are relevant to a

particular study. Yet many education leaders have only a cursory background in research design and statistics. According to a national survey from a recent study, only 36% of education leaders possessed or were working to obtain a research-focused PhD or master's degree (Authors et al., 2017).

To investigate this issue, we developed survey items to measure a construct we defined as Knowledge of Causal Research Design. Collectively, these items were intended to capture education leaders' knowledge of key ideas from statistics and research design, ideas that are critical when evaluating the quality of studies advancing causal claims. The overarching questions that guided this study are as follows:

1. To what extent can we validly measure the Knowledge of Causal Research Design possessed by education leaders?
2. To what degree do education leaders appear to understand critical ideas from statistics and research design?

With these questions as motivation, in this paper we describe the development and use of a set of Knowledge of Causal Research Design survey items with a national sample of education leaders. We also describe two efforts to contextualize the information generated from these items and samples.

Background

The use of research evidence in decision-making is a multi-faceted process. That process typically includes, according to scholars, the acquisition of research, individual and organizational sense-making around research findings, and the incorporation of those findings into practice (Finnigan, Daly, & Che 2013; Honig & Coburn, 2008; Weiss & Bucuvalas, 1980).

Scholars rarely, however, examine education practitioners' ability to evaluate research quality, even though this ability logically interacts with sense-making and use. For instance, we could not find any prior literature that examined how accurately practitioners evaluate the strengths and weaknesses of specific research studies.

To some extent, the federal *WWC* website is intended to alleviate the need for practitioners to evaluate research quality, in that the *WWC* aggregates the findings from experimental studies of specific programs into an easily interpretable "improvement index" and provides information about the strength of evidence related to effectiveness for each program. In some cases, however, practitioners may find themselves in the position of needing to read and understand one or more of the individual studies listed in *WWC* reports, or to investigate topics and interventions or recently published research not covered by *WWC* reviews. Furthermore, we know that many practitioners seek out research directly from colleagues at professional conferences and meetings (Authors et al., 2017). Whether practitioners understand how to evaluate the quality of this research becomes, in such situations, an important question.

What, exactly, practitioners *should* know when reading research is open to debate, and could include some of the content typically found in undergraduate, masters, or even graduate-level courses in statistics and program evaluation. Here again the *WWC* standards are helpful, suggesting several core research design and statistical concepts that can be important when evaluating studies. One set of concepts pertains to whether causal inferences can be drawn from a particular study design, and the extent to which a study's design is susceptible to threats to internal validity such as selection bias or differential attrition (Shadish, Cook & Campbell, 2002). Another key concept is the metric used for reporting effect sizes in many studies, standard deviation units. Researchers often convert estimated causal effects from raw or scale scores into

standard deviation units to facilitate comparisons across studies, yet these "effect size" units may be misinterpreted by many practitioners, for they are not intuitive. Other core topics implicit to understanding *WWC* reports include appropriate sample sizes for testing hypotheses of statistical significance, problems that stem from sample attrition, and the importance of baseline equivalence when making comparisons between treatment and control groups.

For qualitative studies, a similar set of concepts can help practitioners interpret research. For instance, tentative inferences can be made to theory or to other individuals and settings, even from a single case (e.g., Glaser & Strauss, 1967; Lincoln & Guba, 1985), and sampling often proceeds based on the desire to gather contrasting levels of a hypothesized independent variable (King, Keohane & Verba, 1994). However, we know little about what practitioners know about these concepts.

The limited research that does exist in this area provides indirect evidence that practitioners value research quality, but also that few may be proficient in understanding it. Weiss and Bucuvalas (1980) asked practitioners to evaluate research abstracts based on quality, conformity with prior beliefs, their implication for practice, and the extent to which they challenge the status quo. They found that perceived research quality was the strongest predictor of practitioners' sense that the research would be used. In interviews, respondents expressed concerns about the research studies' sample sizes, choice of variables, and research design (pp. 106-107). However, Weiss and Buculavas did not attempt to assess the accuracy of practitioners' evaluations of research quality.

Several recent case studies have examined practitioners' use of both research evidence and data (e.g., student test scores) for the purpose of instructional improvement, and those studies also provide insight into practitioners' understanding of evidence. Farley-Ripple and

Cho (2013, p. 21) observed that policymakers in one district mostly read reports with descriptive data or summaries of research, eschewing both primary research and statistical analyses of their own data. Finnigan, Daly and Chi (2012) noted a similar result, saying "Our qualitative data suggests that even at times when staff used evidence instrumentally (i.e., to inform decisions), they did so in somewhat superficial ways, as they did not appear to have a complex understanding of the various types of evidence, nor did they appear to have the capacity to interpret this evidence in ways that would help them develop appropriate solutions" (p. 14). Other authors suggest that practitioners in districts and schools may not have the necessary expertise to analyze data (Mandinach & Gummer, 2013), though they do not present direct evidence on this count. In fact, most of the studies located in our review present evidence about practitioners that is anecdotal; it does not appear that attempts have been made to create a more formal measure of education practitioners' understanding of topics related to research design and statistics.

There are two related challenges that may help to explain why this issue has not been examined more formally past studies. First, it can be difficult to define and locate an accessible target population of practitioners that will support generalizable inferences. Second, given that practitioners are not licensed or hired on the basis of their understanding of research design and statistics, secondary data on this topic is not readily available. This means that information about this construct must be gathered through the administration of survey items, typically with significant constraints on the time (and effort) busy practitioners are willing to spend answering such items. Along these lines, just as it is critical to have access to a well-defined target population of practitioners, it is equally important to develop items in accordance with a well-defined construct of measurement that can also be administered quickly, to minimize the

demands on practitioners' time. In grappling with both these challenges, this study provides some direct insights about the knowledge of research design and statistics that is held by a specific subset of educational practitioners: district and school leaders responsible for decision-making about educational policies and practices.

Methods

We employed Wilson's (2004) "construct modeling" approach to the development of a set of survey questions (i.e., items) that can be used collectively to measure respondents' skill in evaluating research quality. A construct modeling approach has three iterative stages. In the first stage, one defines, conceptually, a specific measurement construct of interest by creating a "construct map" using prior research and expert opinion; this map defines the construct, attempts to conceptualize it in terms of one or more unidimensional continua, and then delineates steps along the continua that represent salient ordinal distinctions among respondents. In a second stage, experts write items such that the responses to the items will provide evidence about a respondent's location along the construct map. Collectively, scores across a set of items should allow for the sorts of distinctions to be made among respondents that were hypothesized in the initial construct map. In the third stage, data is gathered and item statistics are examined to evaluate whether the items are functioning as expected. If they are, then it becomes possible to use a psychometric model (in this case, item response theory) to relate aggregate item scores back to a location on the hypothesized construct map. The process described above is iterative in the sense that lessons learned during later stages lead to revisions to work done during earlier stages. In rare cases, results may indicate that it is unadvisable to associate a set of items to a

single underlying construct at all, and as we discuss in what follows, the present study represents one of these cases.

Construct Map and Item Development

Our design work began by narrowing and defining our initial conceptualization of the construct of interest, which our team described as both knowledge of basic statistical concepts and of whether specific research designs allow for drawing causal inferences about program impacts. Because of the emphasis on the latter, shorthand label for this construct became "Knowledge of Causal Research Design." The resulting construct map (see Table 1) defined levels of education leaders' hypothesized knowledge about "threats to validity" as described by Shadish, Cook & Campbell (2002). At the highest level, a person would understand many different threats to the validity of causal inferences made in research studies and would be able to understand both the practical and statistical significance of results. At the lowest level of the map, a person evaluating the strength of a causal warrant for a given study would do so primarily on the basis of their intuition and past experiences, and would have no knowledge of statistical terms and concepts. The two levels in between describe education leaders able to recall some mostly isolated definitions and statistical terms (level 2), and education leaders with a deeper conceptual understanding of a variety of key elements of experimental design and how they interrelate (level 3).

Insert Table 1 here

Over a five-month period, we iterated through multiple drafts of items intended to be aligned to the initial construct map shown in Table 1, with the intent that the items would help to differentiate between practitioners at successive levels of the continuum. We ultimately settled on a set of nine items, available upon request, to include on a pilot test. Seven of the nine items were in a traditional multiple-choice format and responses were scored dichotomously as correct or incorrect; the other two items featured nested multiple-choice items, in other words items with one scenario and multiple questions beneath it, such that scores could range from zero to three points. Four of the nine items focused on foundational statistical concepts that might help to distinguish respondents between levels 1 and 2 of the construct map (e.g., how to interpret a percentile, how to interpret a correlation, the difference between the mean and median of a distribution, the definition of an effect size). The other five items posed different short scenarios reflective of different research designs and asked respondents to distinguish between different threats to the validity of inferences that could be made from them. The correct answer in each case corresponded to what should have been viewed as the most significant threat to validity. We split these items onto two forms containing four and five knowledge items each, and administered each form to half our pilot sample.

As we describe in more detail in what follows, the results from our pilot test raised fundamental questions about our premise that Knowledge of Causal Research Design was a measurable construct, at least for our intended target population of education leaders. As a consequence, following the pilot test we switched from an objective of measuring a single construct with multiple items to an objective of accurately characterizing the percentage of national education leaders able to answer purposefully chosen questions about statistical terms and both quantitative and qualitative methods. However, because these items were relatively

time-consuming and could be cognitively taxing for respondents to answer, because the full survey intended to measure a broad set of constructs around knowledge use in a 30-minute instrument (see Authors et al., 2017), and because our pilot test suggested the benefit of having respondents answer a large collection of these might be limited, we administered a random selection of only two to three of these items to each respondent in our national sample.. We considered this a reasonable design strategy since we would no longer expect that the sum or average of our five final "knowledge" items would reliably measure the intended underlying construct.

To select this small set of items, we first identified three items from the pilot test, including one that requires understanding the definition of an effect size ("Effect Size"), another that requires understanding selection bias as the biggest threat to internal validity in the absence of random assignment ("Internal Validity"), and another that requires a respondent to understand why random assignment is important to causal inference ("Random Assignment"). We also added two new items that focused on the ability to evaluate inferences from studies that employ primarily qualitative methods. The first new item presented a scenario in which the results from a case study could be used as the basis for a generalizing to theory ("Generalizing to Theory") and the second focused on the use of purposeful sampling for qualitative case studies ("Purposeful Sampling"). All five items are included in Appendix A.

Pilot and National Samples of Education Leaders

The target population of education leaders consisted of school principals and central office staff from mid- and large-sized U.S. urban districts. We focused on principals and central office leaders because they make the majority of decisions regarding what programs and

interventions schools will adopt. To focus on individuals with instructional decision-making responsibilities, we targeted the following seven roles: deputy, associate and network superintendents; curriculum supervisors; special education supervisors; accountability, assessment, and research coordinators; directors of federal, bilingual, and English as a Second Language (ESL) programs; "multi-role" central office leaders classified in more than one of the above roles; and elementary, middle school, and K–8 principals. We chose K–8 principals because there is more research available on effective programs and interventions at these grade levels than for high school, and because more variety exists in the curricular materials, assessments, and other instructional programs districts may adopt. Because smaller districts may not staff many of the positions included in our sampling frame, we focused on the 1,000 largest U.S. school districts, which according to NCES Common Core data each served more than 9,000 students. We defined strata for our sample with respect to a respondent's role in the district (7 categories) and whether the district was above or below the median enrollment for the 1,000 largest districts (17,860 students). Our accessible target population (i.e., sampling frame) consisted of 41,000 school and district leaders in the 1,000 largest school districts; we purchased the names of districts and leaders from MDR, a firm that maintains a national database of district and school staff. Anticipating a 60% response rate, we generated a stratified random sample of 168 potential survey respondents by role, or 84 for each role-by-size stratum[1].

To assemble a separate sample for the pilot test of our survey items, we excluded education leaders selected for inclusion in our nationally representative sample (as described

[1] We confirmed our rosters and gathered email addresses by searching district websites and contacting districts by phone. If we learned that a sampled individual had left the district or moved to a position not eligible for the survey, we requested name and email information for their replacements. Our survey also included items asking respondents to indicate their role in the district so that we could assess the accuracy of the MDR and project-gathered information about role.

above), and also excluded leaders in the 30 largest school districts in the United States, as these leaders comprised the target sample for a separate study conducted by our research team. After making these exclusions, we drew a random sample of 6,250 individuals from two strata corresponding to leaders who were school principals and those who were not. Because we did not plan to devote resources to follow-up calls and emails, we anticipated a 3.2% response rate and a final sample of about 200 respondents. Our assumption about the low response rate was warranted, as our final pilot test sample was 178, for a response rate of 2.8%.

In contrast, following a series of follow-up requests to participate, our nationally representative sample consisted of 733 individuals with an overall response rate of 51.5%. These individuals came from 487 school districts across 423 cities and 45 states. This represents districts that include roughly 13.8 million of the 50 million students in elementary and secondary students in the United States. Table 2 shows the breakdown of both pilot and national sample respondents by their different professional roles in school districts. In both cases, respondents representing a variety of professional roles participated. Relative to the national sample, our pilot sample contained a slightly higher proportion of curriculum directors and principals, and a slightly lower proportion of deputy/associate superintendents, special education supervisors and people who chose "other" to describe their roles in the school district. The number of respondents who were administered each of the five knowledge items ranged from 263 to 291.

Insert Table 2 here

Results

Findings from Pilot Survey

Insert Table 3 about here

The classical statistics from our pilot test are summarized in Table 3. In general, the items were difficult for education leaders to answer correctly. For the seven items scored dichotomously as correct or incorrect, an average of just 60% of education leaders chose the correct response. Only one item, on the definition for a correlation coefficient, appeared to be easy for respondents to answer correctly (see Figure 1).

Insert Figure 1 here

For the two items for which it was possible to earn either two or three points for a sequence of responses – one focusing on the interpretation of statistical significance, another on threats to internal validity due to sample attrition – the average respondent earned only about half of the maximum possible points.

Notably, the correlations of each item score with the total score of all remaining items on the same form, known as a point-biserial, were extremely low. Point-biserials indicate whether survey respondents who get a particular item correct also tend to do well on the overall test, and vice versa (an incorrect answer correlates with poor overall performance). Higher point-biserials are preferred; a typical rule of thumb in the construction of selected response instruments is to retain items with a point-biserial of about .30 or higher. This rule of thumb might be relaxed in assessments that contain only a small number of items, for a small number of items may translate

to a restricted range of difficulties, which would tend to attenuate point-biserial correlations. However, it is nonetheless notable that only two items had point biserial values greater than .10, and five of the nine items had negative point biserial correlations. Items with negative point-biserials indicate that practitioners who answered the item correctly tended to do worse on the rest of the items on the test relative to practitioners who answered the item incorrectly.

With these near-zero point-biserials, we expected relatively poor test reliability. A reliability coefficient, as estimated using Cronbach's alpha, is a function of two features of an underlying set of item: the total number of items, and the average intercorrelation among the items. In fact, the estimated reliability of total scores on each of our two pilot test forms—the proportion of observed score variability that represents true differences among our respondents—was effectively zero (−.034 and .006). Given the small number of items on each of our two pilot test forms (five and four respectively), we had expected to find a low reliability estimate. Yet a value of zero implies that on average, respondents' answers to any pair of our items were completely uncorrelated. A closer look at the intercorrelations of our pilot items bore this out: only one pair of items had a moderate positive correlation, the rest had correlations that were near zero (or in some instances negative).

Given the results above, we shifted from using multiple items to distinguish individuals' location on our construct map to administering single items, then using data from those single items to characterize the percentage of national education leaders able to answer questions correctly about specific topics related to quantitative and qualitative methods. This effort entailed both administering the items to the national sample as well as conducting additional analysis, including a validity study to ascertain whether the items elicited the knowledge we intended. We discuss each in turn.

Findings from National Survey

Insert Table 4 about here

As shown in Table 4, the results from administering five knowledge items to our national sample appeared consistent with the substantive findings from our pilot sample: education leaders struggled to correctly answer items pertaining to causal research design and analysis. Respondents were most likely to answer our question on purposive sampling in qualitative research correctly (61%); on most other items, half (effect size) or fewer than half (internal validity, random assignment, and generalizing to theory) of the sample answered the item correctly. These results did not vary by the leadership role held by the respondent or by the size of the school district. Results also did not vary by whether the respondent reported that he or she possessed or was working toward a graduate-level degree.

Although we interpret these items one at a time, we can still ask: If we were to regard these five items as a "measure" of something, how reliable would it be? Doing so provides insight into whether our pilot results were an anomaly. Although no person in our sample responded to all five knowledge items, subsets of respondents took different pairwise combinations. As such, we can simulate a correlation matrix (where each cell in the five by five matrix represents a pairwise correlation between two items), and with this correlation matrix in hand, can generate an estimated reliability of a summary score based on all five items. This value, which turns out to be .34, represents the reliability we would expect to observe if the full national sample had responded to all five items, assuming the pairwise correlations are a good estimate of the correlations one would observe if there was no missing data. This can be

interpreted as an indication that only 34% of the variance in an observed sum score based on these five items would be attributable to true variability amongst respondents.

Additional Exploration

We considered two possible explanations for the low correlations among the knowledge items in both our pilot and national survey. The first explanation is that the Knowledge of Causal Research Design construct is only measurable given a precondition that respondents have had some recent and formal exposure to statistics and research design. Without this, one might suspect a vast majority of respondents would be at level 1 of our construct map (recall Table 2), where evaluating strength of research for drawing causal inferences about program impacts is based primarily on intuition. If only a few of our respondents were at level 2, and fewer still were at level 3, then it should come as little surprise that our instrument could not make reliable distinctions among these respondents. A second explanation is that, despite our best efforts, we designed and administered problematic items, in the sense that either the prompt or answer options had equivocal interpretations. In order to explore each explanation, we collected additional data.

Administering knowledge items to a graduate student sample. To start, we reasoned that if the items were well-designed, they should be answered correctly more often when taken by individuals with recent and formal coursework in statistics. With this in mind, we administered the five knowledge items to a convenience sample of 134 graduate students from two universities with Graduate Schools of Education, School A (n=106) and School B (n=25). All of these students had taken at least one masters-level course in statistics, 98% had taken at least two, and 70% had taken at least three.

Unlike the national sample of education leaders, where knowledge items were randomly assigned to subsets of respondents, the graduate student sample answered all five items. Table 4 compares the performance of the graduate student sample on these items to that of the national sample. The graduate students were much more likely than education leaders to correctly answer the three knowledge items that focused on experimental design. Two items that posed questions about internal validity and the benefit of random assignment presented the starkest contrast; 84% and 81%, respectively, of the graduate student sample answered correctly but by only 33% and 46% of the education leader sample answered correctly. The two groups were much more similar in their ability to answer our two qualitative design questions correctly.

We can also use the data from our graduate student sample to examine whether the low reliability found in our pilot test (and projected for the five items administered to our national sample) represented a population-specific finding. To this end, we compared the pairwise correlations between the three experimental-design items administered to all three samples (pilot sample, national sample, graduate sample). These pairwise correlations were considerably stronger for the graduate sample (mean of .21) than for the pilot and national sample (means of −.07 and .05). Through application of the generalized Spearman-Brown formula, which uses

existing data to predict reliability for hypothetical tests of different lengths, it follows that if the graduate sample had been given a set of 20 items with the same average intercorrelation of .21, we would predict that the reliability of the resulting total scores would be .83. In contrast, the predicted reliability for a set of 20 items with an average between-item correlation of .05 (as found for our national sample of education leaders) would be .53.

The graduate student and practitioner samples present an interesting contrast with respect to the two qualitative design and analysis items introduced in the national administration of the survey. Responses to these two items were essentially uncorrelated ($r = -.05$) for the graduate sample but were more strongly correlated for the education leader sample ($r = .15$). As a consequence, the reliability that one would estimate on the basis of all five knowledge items administered to our graduate survey, ($\alpha = .35$), would be virtually identical to the reliability one would project had education leaders answered all five of the same items ($\alpha = .34$).

In summary then, our exploration with the graduate student sample suggests that our experimental design items were sensitive to opportunity to learn, and that responses to them tended to correlate better than they did for practitioners. While there was less difference between the two groups in their ability to answer our two qualitative design and analysis items, the correlation between the items was quite different. When it comes to the experimental design items, reliability appears very much to depend on the opportunity to learn and background of the target population. To some extent, the same can be said when it comes to our qualitative design and analysis items (though it is less clear whether this is due to differences in opportunity to learn). These results support the argument that the lack of reliability observed among scores in our pilot sample was attributable primarily to the characteristics of our target population, and not with the quality of the items per se.

Cognitive interviews with education leaders. We examined the quality of information generated by our knowledge items by conducting cognitive interviews with members of our national sample. We completed interviews with 53 individuals from this sample roughly three months after they completed the original survey, presenting two or three knowledge items to each individual, then asking them to think aloud as they answered the item. We transcribed these interviews, then coded responses according to two criteria: whether the interviewee endorsed the correct or incorrect answer to the item, and whether the interviewee's explanation for the item indicated that they understood or misunderstood the key methodological ideas the item intended to elicit. Thus we coded responses into four categories:

1) Interviewee endorsed the correct answer to the survey item, and provided explicit evidence of understanding the concept that was the intended focus.
2) Interviewee endorsed the correct answer to the survey item, but provided no evidence that they understood the concept that was the intended focus. The latter included clear misconceptions about the concept as well as responses that simply failed to show any evidence of correct thinking.
3) Interviewee endorsed an incorrect answer, but provided evidence of understanding the concept that was the intended focus.
4) Interviewee endorsed an incorrect answer, and gave evidence that they did not understand the concept that was the intended focus. The latter included clear misconceptions about the concept as well as responses that simply failed to show any evidence of correct thinking.

Finally, we also coded "not enough evidence to make an inference" when respondents did not answer the question and when respondents' comments were extremely brief. Table 5 shows

results from the coding scheme described above. Across all items, only 20% of responses had both a correct answer and correct reasoning. Half of responses had an incorrect answer and incorrect reasoning. This means that in roughly 70% of cases, respondents' answer matched the underlying concept of the topic assessed. In another quarter of cases, however, respondents arrived at the correct answer but did not provide evidence that they understood the underlying concept. Only a very small number of responses (2.6%) had correct reasoning but an incorrect answer.

A closer examination suggests that these patterns varied by item. The first item, on effect size, was answered correctly by half the sample, yet none of those respondents' comments suggested that they understood the meaning of this term. Instead, correct answers were arrived at through test-taking strategies and guessing. Similarly, for the generalizing to theory item, all correct responses lacked corresponding evidence of understanding. For the item on random assignment, only half of the correct answers appeared supported by correct reasoning. By contrast, the item on purposeful sampling (see Appendix A) was answered correctly by over half of its respondents, with most engaging in a reasoning process close to what item developers intended:

> So if they want to be able to look at how leadership is really affecting [curriculum implementation] then they should look at schools with a variety of leaders to see if that plays out because if that theory would hold then if you did a purposeful sample with different levels of leadership – you would then see, supposedly, your results playing out in the schools with strong leadership and correlating with not as great results in schools with low leadership.

The same was true for the item on internal validity, which reported on a professional development study whose findings were potentially compromised by selection effects; all correct answers (26%) also noted the possibility of selection effects, often in lay language and often by making connections to their own personal circumstances ("what we deal with the most – is the teachers who participate [in professional development] aren't really the ones that we need to participate. They're the ones who already know it, and they're just getting better.")

Many respondents commented on how far they were from any statistics training ("I'm trying to think back to my college statistics class from 30 years ago"), and many noted that they were not sure of their answers. Many were quick to make personal connections with the content of the research described in each item, and often based their answers on that connection rather than the methodological concept intended. Finally, many respondents went through a process of eliminating responses (i.e., a test-taking strategy) rather than immediately homing in on the correct answer.

Discussion and Conclusion

Taken together, these results indicate that (1) knowledge and skill in evaluating a causal research design may not be measurable within the typical time constraints of a conventional survey instrument and (2) district leaders and decision-makers do not appear to be fluent with some of the foundational concepts necessary to critically evaluate research studies. We discuss both in turn.

Our attempt to use multiple-choice items to measure leaders' knowledge of key statistics and research design principles missed the mark. This occurred during our pilot, in which we used

a construct map to structure our investigation into practitioner knowledge. It also occurred while using single items to characterize the knowledge of a nationally representative sample, where cognitive interviews suggested that for three items, test-taking techniques and guessing often led to correct answers. Low inter-item correlations also suggest that to the extent practitioners do hold this knowledge, they hold it in bits and pieces rather than as part of an organized framework with which they interpret research. Finally, the poor estimated reliability for both the pilot and national samples also may derive from a low average knowledge that restricts the range of true variability among respondents; this will tend to depress estimates of reliability. This suggests that researchers hoping to measure knowledge in this domain with a small set of selected response items may be overly optimistic.

We recommend that future efforts in this area take different form, perhaps beginning with interviews and observations of practitioners interpreting research in local contexts, then moving toward more formal assessments of knowledge, perhaps taking a more creative approach to item format. Revision of the construct may be necessary, or even desirable, based on this qualitative work. Measurement of the original construct, knowledge of causal research design, may also become more feasible if and when education leaders' knowledge increases. Finally, it seems unlikely to us that education leaders' knowledge in this domain can be used with only a few items taken in a few minutes' time; it is more likely that to return reliable scores, an assessment in this domain would take an hour or longer to complete.

Next, our results indicate practitioners' knowledge of causal research design is likely low. Correct answers for individual items hovered around 50%, especially for the national sample. Further, cognitive interviews suggested that when leaders do answer items correctly, guessing

and test-taking techniques were often the cause for three of our five items. This contrasted to much higher levels of performance among graduate students.

Our qualitative analysis also suggests that practitioners may be more likely to hold more intuitive, contextually-bound knowledge in this arena, as described by Level 1 in our construct map. Some correctly reasoned through an item written to elicit knowledge of purposive sampling, connecting their knowledge of how different school leadership contexts affect curriculum implementation to the scenario presented in the item; similarly, practitioners were able to identify voluntary selection into the treatment group as a threat to a non-experimental study based on their own experiences with professional development. By contrast, practitioners who correctly answered the questions on effect size, the benefits of randomization, and generalizing from a single case study were likely to have guessed or used test-taking strategies to do so. We hypothesize that these experiences are more distal from practitioners' experiences, and perhaps more esoteric in nature.

Given these findings, it is not clear how to evaluate the current push for the use of research evidence in decision-making. No Child Left Behind and now ESSA have both emphasized practitioners using research to guide decision-making, yet if practitioners cannot distinguish better from worse experimental design features or interpret effect sizes, that decision-making process may be compromised. The What Works Clearinghouse can act as a scaffold, but as noted above, there are many settings in which practitioners need to evaluate research on their own – when talking to colleagues in local universities and research organizations, when attending a conference, or when finding newly published work. Our findings about causal inference are of particular concern, given that strengthening causal inference in studies of educational interventions has been a focus of federal research dollars for nearly two decades.

Based on these results, policymakers may consider providing easy-to-digest information about research design and statistics to practitioners, for instance in the form of short online courses or accessible rules of thumb. A recent scan of the environment suggests that few such resources exist. Researchers should also be attentive to the challenge of conveying their findings with respect to units that can be meaningfully interpreted, and at a minimum facilitate the comparisons of their results in effect size units to policy-relevant gaps in performance (Lipsey et al., 2012). Finally, institutions that train educational leaders may consider adding coursework targeted toward the practical knowledge leaders need to work with research evidence.

References

Authors, 2017.

Farley-Ripple, E. N., & Cho, V. (2014). Depth of use: How district decision-makers did and did not engage with evidence. In Bowers, A.J., Shoho, A.R., & Barnett, B.G. (Eds). *Using data in schools to inform leadership and decision making*, pp. 229-252.

Finnigan, K. S., Daly, A. J., & Che, J. (2013). Systemwide reform in districts under pressure: The role of social networks in defining, acquiring, using, and diffusing research evidence. *Journal of Educational Administration*, *51*(4), 476-497.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago: Aldine.

Honig, M. I., & Coburn, C. (2008). Evidence-based decision making in school district central offices: Toward a policy and research agenda. *Educational Policy*, *22*(4), 578-608.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic Inquiry. Beverly Hills, CA: Sage Publications.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... & Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *National Center for Special Education Research*.

Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher*, *42*(1), 30-37.

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

Weiss, C. H., & Bucuvalas, M. J. (1980). *Social science research and decision-making*. New York: Columbia University Press.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. London: Routledge.

Table 1
Construct Map: Knowledge of Causal Research Design

4 (critical/evaluative)	Able to critically evaluate research in context Weighs both practical and statistical significance when examining a research report Considers the setting of a research study, and can think through whether a study's conclusions would generalize to a different context Can identify a wide range of potential threats to internal validity
3 (conceptual)	Shows some understanding of what statistical significance actually is, including the importance of sample size ($p = .05$ has a different interpretation when $N = 10{,}000$ relative to $N = 100$.) Has some understanding of why RCTs are good (reducing selection bias), but may not understand that RCTs also have limitations (attrition) Understands fundamental statistical terms such as percentiles, central tendency, variability
2 (recall)	Remembers some terminology, but struggles to explain what it means Knows that randomized control trials (RCTs) are good and that statistical significance is important, but doesn't understand why Recognizes basic concepts such as percentiles and central tendency, but may struggle to understand the relevance of variability
1 (Intuitive)	Understanding of research is based on intuition and "things I've heard"

Table 2
Survey Respondents by Role

	Pilot Survey		National Survey	
Role	Number	Percent	Number	Percent
Deputy, associate and network superintendents	24	14	142	19
Curriculum Supervisors	37	21	73	10
Special Education Supervisors	13	7	81	11
Accountability & Assessment Coordinators	10	6	63	9
Elementary, middle school, & K-8 Principals	41	23	123	17
Directors of federal, bilingual, and ESL programs	8	5	37	5
Multiple Roles	32	18	129	18
Other	12	7	85	12
Total	177	100	733	100

Table 3
Item Statistics from Pilot Test Sample

Item Topic	N	Max Score	Mean	Mean/Max	Pt-Biserial
Correlation	62	1	0.87	0.87	0.16
Random Assignment	62	1	0.57	0.57	0.02
Mean vs. Median	62	1	0.61	0.61	0.08
Statistical Significance^{+}	62	3	1.51	0.50	-0.05
Attrition 1^{+}	62	2	1.01	0.51	-0.01
Attrition 2	100	1	0.52	0.52	-0.25
Control Group	100	1	0.67	0.67	-0.05
Selection Bias	100	1	0.36	0.36	0.16
Effect Size	100	1	0.56	0.56	-0.09

Notes: The first five items in the table above were included in Form A; the next four were in Form B. The respective reliability estimates for the total scores from the two forms was -.034 and .006. $^{+}$ denotes a nested item.

Table 4

Results from Administering Knowledge Items to Samples of Education Leaders and Graduate Students

Item Concept	Education Leaders		Graduate Students	
	N	Proportion Correct	N	Proportion Correct
Effect Size	263	0.52	131	0.66
Internal Validity	291	0.33	132	0.84
Random Assignment	287	0.46	133	0.81
Generalizing to Theory	281	0.28	132	0.36
Purposeful Sampling	263	0.61	132	0.53

Table 5
Results from Cognitive Interviews with 53 Education Leaders from National Sample

Survey Item	Total Responses	Correct, evidence of understanding	Correct, no evidence of understanding	Incorrect, evidence of understanding	Incorrect, no evidence of understanding	Cannot code
Effect Size	22	0 (0.0%)	11 (50.0%)	0 (0.0%)	10 (45.5%)	1 (4.5%)
Internal validity	19	5 (26.3%)	0 (0.0%)	0 (0.0%)	12 (63.2%)	2 (10.5%)
Random Assignment	28	8 (28.6%)	8 (28.6%)	3 (10.7%)	8 (28.6%)	1 (3.6%)
Generalizing to Theory	28	0 (0.0%)	9 (32.1%)	0 (0.0%)	19 (67.9%)	0 (0.0%)
Purposeful sampling	20	11 (55.0%)	1 (5.0%)	0 (0.0%)	8 (40.0%)	0 (0.0%)
All knowledge items	117	24 (20.5%)	29 (24.8%)	3 (2.6%)	57 (48.7%)	4 (3.4%)

Correlation

If the correlation between standardized test scores and family income were strong and positive, we could conclude that:

A. Low family income causes poor performance on standardized tests.
B. On average, students from high-income families score higher on standardized tests than students from low-income families. [Correct Response, 87%]
C. Higher-income families prioritize academic performance more than low-income families.
D. If a family experiences an increase in income, test scores of students in that family will rise.

Figure 1

Appendix A: Knowledge of Quantitative and Qualitative Research Design Items Administered to Samples of Education Leaders and Graduate Students

Effect Size

A large number of students were randomly assigned either to a treatment group that received an intensive tutoring program in reading or to a control group that did not. After participating in the program for 10 weeks, students were given a reading assessment. Results show that students in the treatment group scored higher than students in the control group, with an effect size of 0.3. In this context, what does "an effect size of 0.3" mean?

A. On average, students in the treatment group scored 0.3 percent higher than students in the control group.
B. On average, students in the treatment group scored 0.3 points higher than students in the control group. [correct answer]
C. On average, students in the treatment group scored 0.3 standard deviations higher than students in the control group.
D. The correlation between the curriculum and test scores was 0.3.

Internal Validity

To evaluate the impact of a new summer supplemental math program, a district randomly assigns a large number of students to either a treatment group (which receives written materials to use at home) or a control group (which does not). Which of the following is the biggest threat to the district's ability to draw conclusions based on this study?

A. Students may leave the study after initial assignment.
B. There may be selection bias in the initial assignments to treatment and control groups. [correct answer]
C. Other education interventions may occur in the district during the study.
D. Some students might spend more time using the supplemental math program than other students.

Random Assignment

Imagine that a large district wants to evaluate the impact of a new curriculum. Which of the following is the biggest advantage to randomly assigning 200 teachers (e.g., using a lottery) to either a treatment group (which receives the new curriculum) or a control group (which does not)?

A. Randomization increases the likelihood that the two groups of teachers will be similar in all ways except exposure to the new curriculum. [correct answer]
B. Randomization increases the likelihood that the results of the study will apply to other school districts.
C. Randomization increases the likelihood that the results of the study will be statistically significant.
D. Randomization increases the likelihood that there will be a large difference in outcomes between the treatment and control groups.

Generalizing to Theory

Researchers studied one elementary school teacher's efforts to change her teaching in mathematics and English Language Arts (ELA) in response to new state standards. In ELA, she sought out and actively participated in professional development, asked for advice from colleagues, and created opportunities for collaboration around ELA instruction at her school. In mathematics, she relied exclusively on required professional development workshops and focused on memorizing the material presented so she could apply it in her classroom. Which of the following inferences can you draw from this case?

A. Elementary school teachers' learning experiences may differ depending upon the school subject, and this accounts for why elementary teachers often excel in teaching one subject but not another.
B. Nothing, because the study only involves one teacher.
C. Elementary school teachers' type of engagement in learning may differ by school subject, and these differences may contribute to very different opportunities to learn for teachers, depending on the school subject. [correct answer]
D. Elementary teachers typically change their ELA teaching more easily than their mathematics teaching in response to reform initiatives.

Purposeful Sampling

Researchers randomly sampled six middle schools in order to study the implementation of a new middle school science curriculum. They observed and interviewed teachers over the first three years of using the curriculum. They found that teachers who implemented the curriculum with fidelity worked in schools where leaders learned about the curriculum and allocated time for teachers to talk with one another about it. They concluded that school leadership for instruction was essential for helping teachers to implement the new curriculum with high fidelity. The

researchers have funds to continue the study in six more schools. What would be the best way to provide better support for this conclusion?

A. Randomly sample six more schools to ensure that they can generalize appropriately to the population of middle schools.
B. Select schools not implementing the new science curriculum to provide a comparison group.
C. Select a purposeful sample of schools with different levels of leadership for instruction. [correct answer]
D. The study does not need to be improved, as the design is already rigorous.
E. Collect survey data to standardize the kinds of questions asked of teachers during the study.