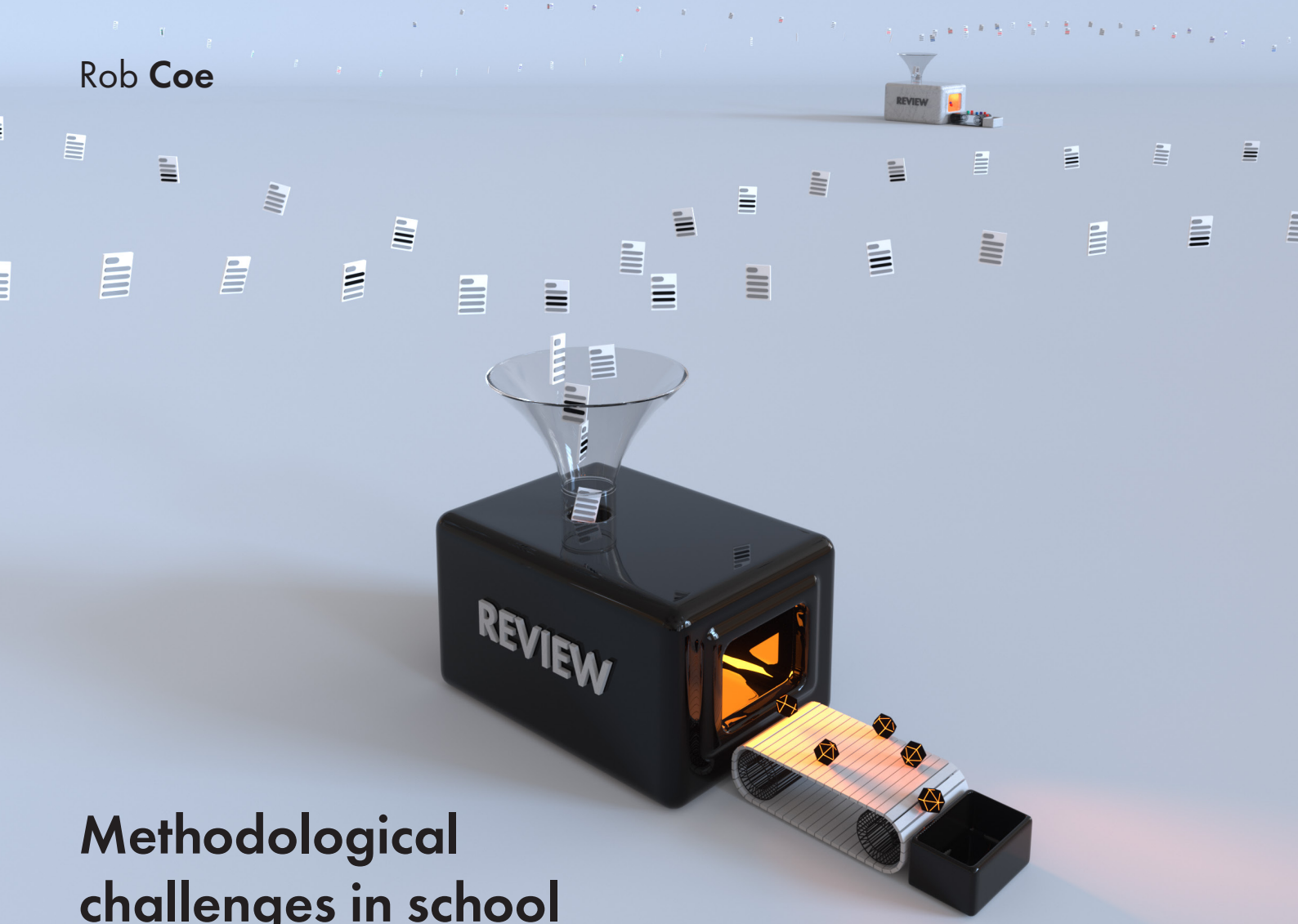
Evidence Based Education

Great Teaching Toolkit

School Environment & Leadership: Evidence Review

Rob **Coe**



Methodological challenges in school leadership research

In partnership with

Cambridge Assessment International Education

The suggested citation for this document is

Coe, R. (2022). Methodological challenges in school leadership research (School environment and leadership: Evidence review). Evidence Based Education.

https://evidencebased.education/school-environment-and-leadership-evidence-review/

Published by Evidence Based Education
in partnership with
Cambridge Assessment International Education

IGNIO
Innovation by design

Contents

Introduction

This paper is part of a series of four that together comprise the *Great Teaching Toolkit School Environment and Leadership Evidence Review*. In order to cater for different audiences, we have split the findings from our evidence review into four separate, but inter-related, documents, of which this is the second.

The first, written primarily for practitioners, and intended to have a constructive, action focus, sets out our *Model for School Environment and Leadership*–the school-level factors that can inhibit or enhance the classroom interactions that promote effective learning. The second (this document) explains in technical detail the key methodological problems faced by research in school leadership, and hence why we are sceptical of many of its claims. The third identifies a selection of studies that we believe contain the most defensible claims and the strongest evidence about the factors we have included in the *Model for School Environment and Leadership*. The fourth provides technical details of the literature search and synthesis process that underpins the other three.

You can find links to download all four papers **here**.

Methodological challenges in school leadership research

Research on school leadership is abundant and influential. A number of key reports have over 1000 citations on Google Scholar (e.g., Elmore, 2000; Fullan, 2003; Leithwood & Jantzi, 2008; Marzano et al., 2005). Claims about the characteristics or behaviours of effective leaders, along with streams of advice for leaders, are likewise abundant. Training in school leadership is also big business, supported by substantial government and private funding in many jurisdictions. Amid all this activity, it seems appropriate to ask about the strength of the evidence base that underpins it: How much do we really know about what makes a great school leader? What do we know about how to help school leaders to be even better?

Unfortunately, when we look critically at this evidence, it seems much of the apparent clarity and confidence disappears. School leaders are presented with plenty of advice, but much of it is not specific enough to be able to follow, its success is dependent on other (not always stated) assumptions or conditions, or it is just not appropriate to their context. Where specific, feasible, appropriate actions can be identified, there seldom seems to be strong causal evidence of likely benefits—pretty much all the research in this area is correlational and descriptive (Liebowitz & Porter, 2019). Many of the widely used terms (for example, instructional leadership, culture, climate) have not yet reached the stage of having widely agreed, common definitions and operationalisations, let alone strong evidence of the validity of the common interpretations of standard instruments. While theory is abundant, robust testing of the predictions that theory makes against independent empirical data seems a lot less common. In short, it is far from clear whether any of this advice is either scientifically trustworthy or practically useful.

This paper sets out to explicate and exemplify some of the main challenges that arise in producing valid and useful knowledge about school leadership. Specifically, we outline three types of problems that can be found in much of the literature on school leadership:

- Constructs that are inadequately defined or measured
- Advice that is too vague to be meaningful or actionable
- Causal claims that are unwarranted

Each of these problems will be elaborated and discussed, using specific examples. The intention is not to focus criticism on particular researchers or studies, but simply to demonstrate that these problems are real. In order to make the point, we have tried to choose examples from prominent researchers in the field and from papers that have been widely cited. The aim is to demonstrate that these problems are endemic and widely ignored by leading researchers in school leadership. Choosing high-status publications for critique also avoids the problem of appearing to publicly attack junior researchers, who may be less robust.

The selection of studies for this review came from a systematic process of identifying relevant studies, and extracting and synthesising their main claims and methods. A full description of the methods used in the search, screening, extraction and synthesis can be found in **Methodology underpinning the Evidence Review**.

Constructs that are inadequately defined or measured

It is a fundamental requirement for scientific enquiry and the accumulation of worthwhile knowledge that we are able to define and operationalise constructs. If people want to talk about, and conduct research on, 'distributed leadership' or 'organisation management', for example, it is important that we all mean the same thing by those words. In research fields where this basic step has not been achieved, we see the 'jingle-jangle' fallacies, where either the same name means different things, or the same construct has different names.[1]

Where definitions are not clear and constructs not well operationalised, we may see the following types of problems:

- Unwarranted interpretations. Looseness in a definition allows a wide range of interpretations, some of which may be inappropriate.
- Confirmation bias. Vague claims are hard to disprove, and may appear to be supported by a wide range of observations that are consistent with them – even though almost anything would be.
- No accumulation of knowledge. If every different researcher uses words to mean slightly different things, there is no prospect of each study building on others' work.
- No independent testing of predictions. Even where theory is developed and predictions made, the lack of shared operationalisations of the constructs make theory and predictions impossible to test.

By contrast, the process of turning a vague concept, defined descriptively, into a clear and well-standardised process for eliciting it forces us to think more clearly about exactly what it is—and isn't. This is illustrated by Lord Kelvin's well-known words:

> "... when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science." (Kelvin, 1891, p. 80)

1 The origin of the phrase 'jingle-jangle fallacies' seems to be Kelley (1927). Kelley attributes the 'jingle' fallacy (assuming that because the same word has been used the referent must be the same) to Thorndike (1904), who in turn attributes it to 'Professor [Herbert Austin] Aikins', without giving a specific reference. The converse 'jangle fallacy' is coined by Kelley to refer to using 'separate words or expressions' for the same thing.

In the rest of this section, we present illustrative examples of measurement problems, taken from the literature on school leadership and environment. The first three relate to the need for clear definitions and operationalisations of constructs and show that without this clarity we may overestimate levels of agreement, fail to accumulate knowledge and continue to talk at cross-purposes. The fourth is an example of an unvalidated survey that may not mean what it appears to. The fifth is illustrative of a more complete and sophisticated validation process, but still falls short of what is required. Finally, we present an outline of what a strong instrument validation process might look like and the requirements for development of robust theory and the accumulation of knowledge.

Vague definitions make it seem as though different models agree more than they do

A review by Hitt and Tucker (2016) presents an overview of the different conceptual models of leadership. They present three 'prominent frameworks' for categorising leadership practices. The three are the 'Five Essential Supports' of Bryk et al. (2010), the 'Learning-Centered Leadership' model (Hallinger & Murphy, 1985; Murphy, 2005), and Leithwood's (2012) 'Ontario Leadership Framework'. A list of the component domains of each is presented by Hitt and Tucker (2016, fig. 5), and summarised here:

- Essential Supports (ES): leadership for change; ambitious instruction; student-centered learning environment; professional capacity; parent/community ties
- Learning-Centered Leadership (LCL): vision for learning; instructional program, curricular program, assessment program; communities of learning; resource acquisition and use, organizational culture; social advocacy
- Ontario Leadership Framework (OLF): setting directions; managing the instructional program; developing people; redesigning the organization

On the face of it, and interpreting these words in their normal English usage, it is striking how little overlap there seems to be between them. Hitt and Tucker's analysis is that, when interpreted in terms of their underlying meaning and practices, of the 28 domains identified, twelve are common to all three frameworks. They offer a 'unified model' that combines all three under five broad domains, each of which contains a number of dimensions.

For example, Hitt and Tucker's (2016) second domain is succinctly titled 'building professional capacity'. Under this heading they include staffing (sometimes grouped with more administrative/managerial tasks), relational trust (often judged to deserve a category of its own), accountability (which they acknowledge others put elsewhere) and 'values and beliefs about teacher responsibility for change'[2] (elsewhere labelled as 'teacher efficacy', and discussed below).

A more scientific approach to reconciling these different models would be to identify specific areas in which they make different predictions and then design empirical tests to see which stands up best. Unfortunately, the models are all so vague that none of them really makes any kind of testable prediction.

2 Although, somewhat confusingly, 'values and beliefs about responsibility' also appears in domain 3, 'creating a supportive organization for learning'.

Vague definitions prevent the accumulation of knowledge

A feature of the history of the school leadership literature (as presented in reviews, e.g., Hitt & Tucker, 2016; Leithwood et al., 2004) is the emergence of distinct leadership 'styles'. Given that none of these styles seems particularly well-defined, nor is there any empirical evidence that any of them is a good fit to anything observable, nor is there an agreed way of diagnosing which one applies in a particular context, they seem to have little scientific value. The fact that they ring true as a narrative should not distract us here: many things that seem intuitively undeniable are in fact wrong.[3] Popular adjectives in the leadership world include 'instructional leadership', 'transformational leadership' and 'distributed leadership'.

The definitions of 'instructional leadership' are neither clear nor universally agreed (Grissom et al., 2021). Robinson et al. define pedagogical leadership as "close involvement by leadership in establishing an academic mission, monitoring and providing feedback on teaching and learning, and promoting professional development" (2009, p. 88). In other words, it is a collection of at least three things. Hallinger has a slightly different definition, again with three dimensions, but not quite the same three ('defining the school's mission', 'managing the instructional program', and 'promoting a positive school learning climate', 2005, p. 225). The three are subdivided into ten 'instructional leadership functions' which cover a very wide range of activities, not all of which seem obviously 'instructional' (e.g., 'communicating the school's goals' or 'maintaining high visibility').

Despite the lack of clarity about what 'instructional leadership' actually is, there seems to be a lot of agreement that it is a good thing (Barker & Rees, 2020). Interestingly, even the proponents of 'instructional leadership' acknowledge that school leaders typically spend little time doing it. Studies that observe what headteachers actually spend time on (e.g., observational studies, time logs, or self-report) "have tended to yield a similar picture of principals who allocate a relatively small proportion of their time to instructional leadership" (Hallinger & Wang, 2015, p. 9), but instead are drawn "towards managerial and political leadership roles".

3 For examples, see this list of cognitive biases **https://en.wikipedia.org/wiki/List_of_cognitive_biases**

In relation to 'distributed leadership', claims about its power (Leithwood et al., 2019) seem to be unencumbered by the fact that researchers have "not yet reached a consensus on what distributed leadership is"(Tian et al., 2016). As Harris et al. (2007, p. 338) note,

> "Part of the appeal of distributed leadership resides in its chameleon-like quality; it means different things to different people. This is also its central weakness. Distributed leadership has become a convenient way of labeling all forms of shared leadership activity."

Curiously, it is not hard to find acknowledgements of the vagueness of these labels within the leadership literature. For example:

> "The lesson here is that we need to be skeptical about the 'leadership by adjective' literature. Sometimes these adjectives have real meaning, but sometimes they mask the more important underlying themes common to successful leadership, regardless of the style being advocated." (Leithwood et al., 2004)

A systematic review and meta-analysis by Witziers et al. (2003) found no overall significant relationship between leadership practices and student outcomes. Explaining the reasons why different systematic reviews of the same literature find different results is always tricky, but one difference appears to be the use of more rigorous thresholds of measurement quality in their inclusion criteria. It may be that by excluding studies with weaker measurement of either educational leadership or of attainment, Witziers et al. also removed the evidence of an overall association: "Consistency in the way concepts are operationalized is not the strongest feature of leadership research" (2003, p. 406).

Vague definitions allow jingle-jangle confusion

The phrase 'jingle-jangle' has been used to describe the problem of either applying the same label to things that are different with the assumption they are the same, or assuming that different names for the same thing will refer to different things (Kelley, 1927, see note 1, above). An example of a construct that illustrates both problems is 'teacher efficacy'.

The word 'efficacy' draws on Bandura's (1977) theory of self-efficacy, but also has its origin in Rotter's (1966) work on locus of control. Measures of teacher efficacy, from self-report surveys, have been found to predict a range of outcomes, including student attainment and motivation, as well as teacher persistence, resilience and flexibility (Tschannen-Moran et al., 1998). A number of different surveys have been developed to measure teacher efficacy, along with related measures of 'collective efficacy' and 'principal efficacy'.

From consideration of the component items in these surveys, it seems that 'teacher efficacy' draws on:

- teachers' beliefs about the likelihood of good outcomes;
- teachers' assessments of their own professional skill, knowledge and expertise;
- teachers' perceptions of their colleagues' expertise (sometimes called 'collective efficacy');
- teachers' perceptions of the relative importance of different factors (particularly their own agency) in determining student outcomes (locus of control); and
- teachers' willingness to demand high standards from students, and to persist when they are not easily delivered.

Teacher efficacy is acknowledged to comprise two underlying components (Pajares, 1997): teachers' assessments of their own competence, and their expectations that teaching can influence student learning. And it is recognised, particularly in relation to the former component, that the same teacher's self-assessments may vary for different aspects of their role. Nevertheless, when different instruments emphasise different components in their measure of 'teacher efficacy', there is scope for confusion about what it means.

Elsewhere, similar constructs that could certainly fit within this 'efficacy' heading are given different names, for example:

- Bryk et al. (2010) have 'Collective responsibility' which includes 'How many teachers in this school feel responsible that all students learn?';
- Hallinger (2005) has 'Teaching effectiveness' which includes 'When I try really hard, I am able to reach even the most difficult students'.

Unvalidated measures may not mean what we think they do

The example here is taken from a study that used exploratory factor analysis of survey responses to generate and interpret a factor and make claims about its interpretation that have since been widely cited, but without any attempt to validate those interpretations.

The study is by Grissom and Loeb (2011). Their surveys of perceptions (of both principals and assistant principals) of the principal's effectiveness in a range of tasks found that a factor they identify as 'organization management' had the highest correlation with school accountability ratings, teacher satisfaction and parents' rating of the school, after adjusting for a range of school factors, including accountability rating in previous years. This study has been widely cited (it has 726 citations on Google Scholar), including being cited unproblematically in key reviews (e.g., Hitt & Tucker, 2016; Liebowitz & Porter, 2019; Robinson & Gray, 2019), and the construct of 'organization management' has been used in a range of other studies.

Our interest is in what the factor labelled as 'organization management' really means in all these studies. It is not clear that there is a theoretically driven, conceptually coherent and operationally clear definition—certainly not one that is shared by all studies. It often seems, for example, that it is a bit of a catch-all category for things headteachers do that are not directly focused on instruction, curriculum, assessment or professional development.

In Grissom and Loeb's (2011) study there were two measures of 'organization management' within each school, derived from the perceptions of principals and their assistants, respectively. The school-level correlation between them was modest (0.15). A key part of the validation argument to establish whether we can interpret a score as the principal's effectiveness in 'organization management' would be to show that the perceptions of different actors within the same school were in agreement. A correlation as low as 0.15 is pretty unconvincing in this respect. Moreover, although the same items were used in both versions of the survey, the factor loadings were quite different (see Grissom & Loeb, 2009). For the items whose loadings are provided (those above 0.3) and that load highest on 'organization management' in both versions of the survey, the correlation between the two loadings is -0.04. And, of course, a 'factor' that emerges from exploratory factor analysis is a weighted sum of all the items in the survey, not just the ones with loadings above some threshold—which makes it pretty much impossible to interpret anyway.[4] If we really want to make a claim about a measure of 'organization management', we need to develop and validate a measure of it, and then use the same measure consistently.

4 See, for example, Protzko (2022) for a study in which nonsense items generated a scale with high factor loadings.

Among the things Grissom and Loeb might have done to offer more convincing evidence that their score could be interpreted as a measure of a school's organisation management are:

- Provide a conceptional and theoretical rationale for the meaning and importance of the construct;
- Provide a clear definition, including examples, non-examples and boundary cases;
- Identify a set of survey items that operationalise the construct;
- Report the internal consistency of responses to these items;
- Report the intra-cluster correlation for the responses of assistant heads in the same school;
- Report the stability of the measure for the same unit over time;
- Evaluate the extent to which associations between the measure and measures of other theoretically related constructs fit with what the theory predicts; and
- Conduct a full multi-trait multi-method analysis.

Partial validation is better than none, but still not enough

A second example of the need for full validation of the instruments used in leadership research is drawn from Hallinger and Wang (2015), which presents validation evidence about the PIMRS (Principal Instructional Management Rating Scale), developed by Hallinger in the 1980s, and claimed to be "the most widely used instrument for studying principal leadership in the world" (Hallinger & Wang, 2015, p. xv). Yet, in a book-length account of the rationale, development and validation of this instrument, the authors report that "to date, researchers have uniformly focused on assessing reliability of the PIMRS through tests of internal consistency".

Indeed, it is notable that, when any kind of validation evidence is offered, internal consistency is about as far as most researchers in this field seem to get. The findings of McCrae et al. (2011) on the limitations of internal consistency are relevant here: "Internal consistency of scales can be useful as a check on data quality, but appears to be of limited utility for evaluating the potential validity of developed scales, and it should not be used as a substitute for retest reliability ... internal consistency and retest reliability are unrelated variables."

Hallinger and Wang's (2015) book seems to offer a more impressive and full account of the development and validation of a school environment instrument than any other we have found. For example, it grounds the structure of the instrument in the best available research; it describes a detailed process of developing the survey items, taking account of expert views of school leaders and researchers. It also recognises that where teachers rate aspects of their school or principal, the unit of analysis is the school, and discussion of 'reliability' must take account of both the level of agreement among items (i.e., Cronbach's alpha) and the level of agreement among teachers in the same school (i.e., intra-class correlation).[5] Furthermore, it uses Wilson's (2005) 'Four Building Blocks' approach to assessment design and applies the Rasch measurement model in this.[6] It reports item fit and differential item functioning (DIF), and it analyses the relationship between the PIMRS measure and other leadership instruments; it even applies the unforgiving multi-trait multi-method (MTMM) approach. All of this may be seen as best practice in instrument development.

Of course, a critic might question why the author of an instrument that was first developed and advocated in 1985 should wait 30 years to provide this kind of validity evidence. A pedant might further quibble that, after explaining that responses to the teacher survey should be analysed at the school level, Hallinger and Wang (2015) should not then apply their Rasch analysis to individual teacher responses.

5 In fact, Hallinger and Wang (2015) do not mention ICC as such, but refer to Ebel's (1951) test and use an indicator of reliability that combines both aspects using Generalisability Theory.

6 Unfortunately, the reference given to Wilson (2005) is wrong, but the account is recognisable as being from this source.

But there is a more heart-breaking disappointment in store for the reader whose hopes are raised that, at last, here is a model example of validation practice.

At the end of the section in which the authors have justified the MTMM approach and identified the other measures against which they can validate the PIMRS, they say:

> "Our studies of the external validity (i.e., concurrent, convergent, divergent) of the PIMRS are in progress. Thus far, the preliminary pattern of results support the proposition that the PIMRS meets expected standards of concurrent, convergent and divergent validity. However, we will refrain from asserting this claim until the data are ready for peer review." (Hallinger & Wang, 2015, p. 110)

The next section promises "validation through assessment of impact on criterion variables", but turns out to be just a synthesis of correlations "between the PIMRS constructs and student achievement"—so not really 'impact' as we should understand it (see below for more discussion of this). But here again, no actual data are presented:

> "We plan to use a combination of research synthesis and meta-analysis in order to examine the pattern of results across these studies as a means of shedding light on another aspect of the external validity of the PIMRS. We expect to be able to report these results within the next two years." (2015, p. 111)

In searching through the publications of these authors in the six or so years since that was written, and having attempted to contact the author, we have so far been unable to find these analyses.

Recommendations for adequate validation: Minimum requirements for valid interpretation of measures

Some requirements for best practice in measurement include:

Theory-driven design of instruments, in line with best evidence (e.g., Gehlbach & Artino, 2018; Gehlbach & Brinkworth, 2011; Lietz, 2010). For example, Gehlbach and Artino (2018) advise us to "avoid questions with agree–disagree response items, employ questions with construct-specific response options, ask only one question at a time, use positive language, avoid reverse-scored items, and carefully choose item formats to answer the question asked".

Well-specified, replicable instruments. A minimum requirement is that survey items are provided. When they are, the interpretations must then match the meanings of these items.

Acknowledgement of mono-method biases (common-source and common-method). Most measures of school climate or environment use surveys. We should always be cautious about interpreting relationships among constructs that are measured using the same method and respondents, since the common method can introduce a range of spurious response sets that distort these relationships—either to exaggerate or mask (e.g., acquiescence, halo effects or social desirability). Goldring et al. (2008) try to measure principals' knowledge and self-perceptions, using MTMM approach. Unfortunately, the within-trait, cross-method correlations they report are close to zero.

Appropriate unit of measurement/analysis. If the measure is supposed to capture something about a school or its leadership, then analysing individual teachers' responses is not appropriate. It is best to use a multilevel model (e.g., multilevel structural equation model) or analyse at the school level. We need to know the extent to which teachers at the same school share similar perceptions of things like 'culture' or 'leadership style'. If they don't agree, it is hard to see how this can be a characteristic of the school; any relationships among teacher-level variables (e.g., Cronbach's alpha or path coefficients) are just reflecting individual differences among teachers at the same school.

Recognition of overlap/multicollinearity/discriminant validity. Are these dimensions really separate? For example, Liebowitz and Porter (2019) calculate separate 'effects' of principals' instructional management and internal relations on student outcomes, but report that the correlation between them, at study level, is 0.72, and, for effects on school organisational health, 0.98. If the correlation between a subscale and the overall measure is higher than the reliability of the subscale, then the overall score is likely to be a better estimate of the true subscale score than the observed subscale score itself.

Justification of interpretations of instrument scores. A validation process should include at least the following three parts:

- **Content validation**. Clear definition and theoretical rationale for the construct. Presentation of the items. Description of the scoring/analysis process. Ideally, judgements of experts that the items represent the whole of the construct and are not influenced by irrelevant features.
- **Internal validation**. Internal consistency among the items in the scale (i.e., Cronbach's alpha). Ideally, this is calculated using a different sample from the one that led to any grouping or selection of a larger set of items. If there are multiple responses within an analysis unit (e.g., teacher surveys are interpreted as giving a measure of the school) then we also need to know the level of agreement among responses in the same unit (e.g., intra-cluster correlation). Also useful to know would be analysis of item discrimination, information, differential item functioning and scale dimensionality.
- **External validation**. Evidence that the measure correlates with other measures of the same thing (especially with other measures that use a different method) and correlates in expected ways with a range of other things (including some that should be independent: discriminant validity).

Independent replication of validation processes. As Grissom et al. (2021) note, "we observed dizzying variation in what factors leadership studies considered, how those factors were operationalized, and the approaches the studies employed for analyses. Even among studies of the same topic, we seldom encountered two studies using the same measurement tools, or studies that replicated an earlier result."

Advice that is too vague to be meaningful or actionable

The school leadership literature contains plenty of advice for school leaders. Its recommendations seem intuitively compelling—indeed, impossible to disagree with—but ultimately lacking in specificity. It is often reminiscent of the kinds of advice found in astrology or in common proverbs. For example, we might all agree that 'many hands make light work' and also that 'too many cooks spoil the broth', even though they are essentially contradictory. If we hear either alone, we can easily think of contexts in which it seems true. More practically useful, however, would be to understand the conditions under which having more people to do a job is better, and when it is worse. Such advice uses the combination of vagueness and superficially obvious truth, usually framed to seem positive, that leads them to be universally endorsed, and has sometimes been described as a 'Forer' or 'Barnum' effect (Forer, 1949; Meehl, 1956).

- Instead, for a piece of advice to have practical value and be trustworthy, we would need:
- Clarity. Advice is operationally clear: there would be no doubt about whether a particular behaviour in a particular context was in fact an example of following the advice.
- Falsifiable. We must be able to specify conditions in which an observation would refute the advice.
- Malleable behaviours. It must advise us to do something we could change by deliberate choice.
- Viable alternative. There must be an actual choice: instead of doing the advised thing, we might plausibly do something else.
- Evaluation evidence. Robust comparisons of the impact of following each choice, using a design that can support causal claims (e.g., RCTs), reporting impact on well-measured outcomes that matter.
- Theory. Some level of understanding of the relevant mechanisms, support factors, moderators, etc., that determine the conditions under which each choice is best.

To illustrate the difficulty, we can start with the first of the 'specific leadership practices' on Leithwood et al.'s (2019) list, to 'build a shared vision'. This is more of a description of a desired end point than a recommendation for action, but there is a clear implication in the whole list that these are things a leader should try to do.

There are many different ways a school leader could set about trying to achieve it, guided by quite different sets of underlying principles. For example, the following three approaches, all of which are unpalatable caricatures, are all compatible with the advice to 'build a shared vision':

1. 'Top-down': From the outset, and repeatedly, leaders should enunciate their own long-term aims for what the school should be like, in the hope that if staff hear it often enough they will come to accept it (or at least stop arguing against it).
2. 'Democratic': Leaders should facilitate a process of getting all their colleagues and other stakeholders to exchange their views about what the school should be like and trying to identify and amplify any common ground to build a consensus on which (after endless time spent debating) all can agree.
3. 'Authoritarian': Leaders should clarify and express their values and beliefs about what the school should be like, and manage out or silence any colleagues who do not agree with them.

In relation to the advice to 'build a shared vision', none of the six bullet point requirements above has been met. It is also worth noting that a list of 19 things that 'successful school leaders do' is probably too many to be useful. An open-minded school leader who is looking for guidance in the literature about what they should do can probably focus in any serious way on at most one or two of these at a time. Where should they start?

What we need is:

- Specification of leaders' alterable characteristics (e.g., behaviours, knowledge, skills, beliefs) that are operationally well-defined;
- Good evidence about how these characteristics can be altered; and
- Good evidence that improving these leader characteristics leads to improvements in valued outcomes.

A similar lack of clarity can be found in the attempts to describe the kinds of characteristics and abilities of leaders that comprise good leadership. Again, we can take an example from Leithwood et al.'s (2019) summary of the state of the art on school leadership research, and select 'problem-solving expertise', their first example of 'personal leadership resources', to illustrate.

'Problem-solving expertise' is a characteristic for which anyone who hears it will have a sense of what it means, so again, it appeals to our intuitive sense of truthiness. Moreover, no one is going to argue that expertise in solving problems is irrelevant for school leaders. The difficulty, if we want this claim to stand up to any kind of scientific scrutiny, is that we may all have slightly different conceptions of what it means.

Without even considering the 'solving' aspect of problem-solving, if we just focus on the types of problems we might have in mind, there may be quite a range. For example, any of these are problems that a school leader might have to resolve:

- Two members of staff have a dispute and refuse to work together.
- A dog runs into the playground and creates havoc.
- Some students' ability to comprehend written text seems to be below the level required for them to access the curriculum fully.
- Three teachers call in sick on a morning and none of the usual supply teachers are available.
- The school's IT system crashes.
- The judgements of senior staff, based on lesson observations, are that a particular teacher's classroom practice is weak.
- One of your teachers makes a comment that is interpreted as blasphemous and offensive by some members of the school community.
- The curriculum seems to be too full of content – teachers are unable to get all students to a level of mastery of it.
- A group of young people are selling drugs outside the school gates.

The point is that a school leader who copes well with solving one of these types of problems may not be so good with another. And, of course, we can also ask how the context in which the problem is manifested interacts with the nature of the problem and its solutions, what criteria we use for evaluating 'solutions', and whether the 'expertise' is a characteristic of an individual leader that is stable and generalisable? Until we have addressed all these issues, we have not really defined what we mean by 'problem-solving expertise'.

Causal claims that are unwarranted

It is a truism that correlation does not imply causation. Indeed, this is sometimes explicitly acknowledged in writing about school leadership. Unfortunately, however, even some writers who do acknowledge it—as well as others who fail to do so—also make claims that are essentially causal in nature. Given the level of cognitive dissonance that would be caused by writing in the same article that causal claims are not warranted and then also making such claims, it seems likely that the writers do not believe they have made a causal claim. So it is important to clarify what we mean by this.

Sometimes a causal claim is explicit and obvious; more often, it is implicit. If a writer uses words like 'effect', 'affect', 'impact', 'leads to', 'benefits from', 'influences', an implicit causal claim has been made. A key test is whether the variables could be interchanged: a correlation is symmetric and it means the same to say 'A is associated with B' as the other way round. None of these words can be interchanged in that way.

The word 'effects' is perhaps more ambiguous because it is conventionally used to refer to the coefficients in a regression model, even where there is no explicit argument about why a causal interpretation is appropriate and such an interpretation may not even be intended. For example, we may talk about 'fixed effects models' without thinking causation is implied. This convention is unfortunate, but well established.

"As is common in the school effectiveness literature, we use the term school effects to indicate statistically significant associations between variables. These associations do not need to be causal in nature" (Hallinger & Wang, 2015, p. 35).

At other times, a causal claim is implied by the authors putting forward a policy recommendation. It cannot make sense to suggest making a change in A as a possible way to improve B unless you think A influences (causes) B.

Examples of unwarranted causal claims

Our first example comes from Leithwood and Jantzi. Somewhat puzzlingly, they state: "Our data do not permit us to make strong claims about cause-and-effect relationships. Nonetheless, we use the language of effects throughout as an indication of the nature of the relationships in which we were interested" (2008, p. 514). Ideally, one might hope that a researcher who was interested in causal relationships would forgo the temptation to use language that implied such relationships unless justified by the kind of data they actually did have—even where that might be inconvenient. Although it is welcome to see researchers explicitly stating a limitation of their research design, the virtue of stating it is not sufficient to grant them licence to then ignore it.

Leithwood and Jantzi (2008) focus on leaders' individual and collective efficacy and report various correlations between different constructs taken from surveys, one for teachers, one for principals, from 96 schools. For student achievement they have attainment measures of '% proficient' at school level each year (2003-5), and a change measure (difference between '% proficient' in 2005 and 2003). Principals' self-reports of their efficacy do correlate positively with the proficiency percentages (mean r for combined measure is 0.26), but not with the change score (r = 0.05).[7]

There are a number of examples in this paper of the kinds of implicit causal claim described above. For example, a sentence like "[o]ur own study examined the influence of leader efficacy on leader behavior, on the school and classroom conditions that we judged to have the greatest impact on student learning ... and on student learning itself" (Leithwood & Jantzi, 2008, p. 509) clearly implies an asymmetric relationship; the names of the variables cannot be interchanged without changing the meaning. The word 'effects' is used several times in the context of regression models (e.g., "Standard regression equations were used to estimate the effects of LSE, LCE and an aggregate measure of efficacy on leader behavior as well as school and classroom conditions", p. 517). They also characterise the results of their structural equation model as showing "the causes and consequences of school leaders' efficacy beliefs" (p. 519). The word 'effect' is used repeatedly in the conclusion; for example, "There was a stronger, though still moderate, effect of aggregate leader efficacy on both classroom and (especially) school conditions" (p. 522). If the intended meaning is that these are symmetric associations in which the variables could be interchanged without changing the meaning, it seems at best poorly communicated, and at worst misleading.

An interpretation entirely consistent with the evidence presented by Leithwood and Jantzi (2008) would be that in schools with higher-attaining populations of students, principals report their perceptions of the capacity of themselves and their staff to promote good outcomes slightly more positively than in lower-attaining schools.

7 No confidence intervals are given, but Fisher's (1921) method with 96 pairs gives a 95%CI for 0.05 of [-0.15, 0.25]

This might be partly explained as a self-serving attribution bias (Bradley, 1978; Kennedy, 2010; Wang & Hall, 2018); people are generally more likely to attribute positive things to their own abilities or actions, and negative things to factors outside their control. It may also reflect the fact that some schools genuinely face greater challenges than others: schools serving the most disadvantaged communities may rightly perceive their efficacy as lower. Crucially, the lack of any relationship between these perceptions and a school's improvement trajectory—and the fact that the perception surveys were completed at the end of the period for which attainment data were reported—makes it hard to justify the claim, on the basis of this evidence, that efficacy perceptions are a causal driver of improvement.

A second example of the uncritical assumption that correlations should be interpreted in the obvious way can be found in the study by Grissom and Loeb (2011). Again, this is a correlational study, albeit with a reasonably full list of explanatory variables, for which they control. The main finding is that their measure of 'organization management' (as discussed above) is correlated with three school outcomes: accountability ratings, teacher satisfaction and parents' rating of the school, after adjusting for a range of factors. In principle, if a correlational study controls for all possible confounds, it may be warranted to interpret the relationship as causal. Such a claim is precarious, however, as a critic has only to suggest a single, plausible, omitted variable that could account for the link for it to falter. In this case, we are happy to accept that they have sufficient controls to shift the claim from 'organization management is correlated with school accountability measure' to a claim something like 'organization management is correlated with the school's impact on learning'.

In their discussion section, Grissom and Loeb claim: "the results suggest that reallocating principals with higher organization management competencies to schools with larger numbers of high-poverty students could be a meaningful way to address socioeconomic achievement gaps" (2011, p. 118). This seems to imply that they think this is a causal mechanism, and perhaps even one that could be manipulated to influence outcomes. This implication is strengthened by additional recommendations for policy that both appointment processes and programmes for the preparation and professional development of principals should incorporate a focus on their organization management skills.

However, the results reported by Grissom and Loeb (2011) are entirely compatible with an explanation in which the causality is reversed. For example, if principals in more effective schools have more time available for organization management, or benefit from some kind of 'halo' effect that enhances both their self-perceptions and those of their vice-principals, we might expect to see the same correlations. Grissom and Loeb do not appear to consider either possibility.

Evaluations of leadership interventions

The kind of study that could provide strong evidence to support a causal claim would be one in which a well-specified intervention is robustly evaluated. In the field of school leadership there are a small number of intervention studies, so we should consider whether they provide a basis for warranted causal claims.

In conducting their systematic review of evidence about the effects of principals, Grissom et al. (2021) find ten evaluation studies of eight interventions. These programmes contain a mixture of training, induction and in-role support, including professional learning communities, coaching, observation and structured feedback. Many also involve competitive selection on to the programme, though others are aimed at principals of under-performing schools. Some interventions are targeted at early-career or new principals, while others are open to all. In reviewing these evaluations, Grissom et al. view this latter distinction as important and conclude that programmes to support early-career heads "generally showed evidence of positive effects on achievement", while "interventions to support a mix of early-career and other principals are less likely to show evidence of positive impacts on student learning" (2021, p. 79).

However, there may be other explanations for this apparent difference, given the mix of different contexts, content of the interventions, and evaluation designs in the different programmes and their evaluations. For example, if programmes targeted at new leaders tend to attract ambitious, high-achieving volunteers, while those for existing principals often end up being used as required remediation for conscripts whose schools are not doing well, it probably would not be appropriate to attribute any differences in their subsequent schools' performance to the programmes.

Another confound is that only two of the evaluations (Herman et al., 2017; Jacob et al., 2015) used random allocation to control for selection effects and create a genuinely comparable comparison group. Both were evaluations of programmes for a mix of new and established principals; both found no effects. Perhaps relevant to the question (discussed above) of whether high teacher self-efficacy is the cause of increased attainment is the finding from Jacob et al. (2015) that their programme did lead to improvements in participants' self-efficacy. This seems to be a significant challenge to the view that efficacy perceptions are a causal lever to raise student attainment.

For the non-randomised designs, we must consider the possibility of unobserved differences between programme participants and others as a threat to the interpretation that the programme was the cause of any differences between their schools at the end.

For example, in Gates et al.'s evaluation of *New Leaders' Aspiring Principals Program*, the description of the selection process makes it seem quite intensive: applicants are assessed on their interest, knowledge of pedagogy, communication skills, interpersonal relationships, data-driven decision-making—among other things (2019, p. 9). There were also ongoing assessments throughout the programme, informing a decision to endorse the candidates for principal placement at the end (p. 22). We are not told what proportion of eligible principals apply, or what proportion of applicants are accepted onto, or graduate from, the programme, but it seems likely that those who do may have been initially quite different in important ways from leaders of other schools in the same district. Even if the observed differences in student outcomes can be attributed to the impact of the school principal, it is not clear to what extent the New Leaders program has added to this.

Indeed, Gates et al. point out that "one must be careful in attributing characteristics that are observed in program graduates to the program activities alone" (2019, p. 19). They also note that a range of other unobserved factors, for example any differences in the trajectories of the schools, their hiring policies, or other "characteristics of the neighborhood or resources available to the school", could also bias the results, and hence acknowledge the limitations in their design: "our evidence of causal effects should be interpreted with some caution" (p. 36). Crucially, having stated this caveat they do not then ignore it, but present their results as associations rather than causal effects. Unfortunately, the use of the word 'effects' in the review by Grissom et al. (2021) somewhat misrepresents this.

In addition to the two RCTs mentioned above, there are a handful of others. Jacob et al. (2015) were able to locate two RCTs prior to 1987, but only one since then, by Goldring et al. (2008)—a small study with high attrition that found no evidence of impact. A substantial, systematic review of school leadership improvement interventions by Herman et al. (2017) used the Every Student Succeeds Act's (ESSA) Evidence Tiers to categorise the weight of evidence for impact. Tier I (Strong evidence) requires "at least one well-designed and well-implemented experimental study (randomized controlled trial)". Under Tier I they found only the Jacob et al. (2015) evaluation that found no impact on student achievement and three evaluations of the Knowledge Is Power Program (KIPP)—a whole-school improvement model that includes a focus on 'power to lead' as one of its five pillars; all three KIPP studies found positive impacts on student achievement. However, these evaluations of KIPP schools (where students are randomly allocated to schools in lotteries; there is no random allocation of principals) do not really tell us much about the impact of leadership interventions per se, given all the other components of KIPP. Although the KIPP approach depends heavily on school leaders, who receive very substantial training and support, it is highly selective (just 7% of principal applicants receive offers to train) and KIPP leaders are different from other school leaders in a range of ways that predate their training (Macey et al., 2009).

While it may be true that students who attend KIPP schools do better than they would have done at other schools, it is impossible to say how much of that benefit is attributable to any differences in the leadership practices or leader characteristics of those schools, or to the training and support their leaders receive.

In their review of leadership research, Liebowitz and Porter conclude that "the contribution of principal training programs to principals' influence on student outcomes appears to be modest at best and presents measurement challenges" (2019, p. 3). In this summary they include the evaluations by Gates et al. (2019) of the *New Leaders* programme (discussed above) among those with positive effects, as well as others with mixed or apparently negative effects.

Overall, therefore, the evidence that any kind of training or support programme for school leaders can be said to improve student outcomes is pretty unconvincing at present. We simply cannot yet claim that we know how to help leaders to be more effective.

Conclusion

We have argued that a significant part of the school leadership literature, including some of the most cited and influential studies, contains methodological flaws that make its claims untrustworthy. Common practice in measurement falls short of acceptable standards: in defining its concepts, designing and constructing measures to operationalise them, and validating the interpretation of those measures. Much of the plentiful and eagerly attended advice that is given is not clear, actionable or scientifically verified. And the field is riddled with spurious, often implicit, causal claims that are simply not warranted.

Some of these limitations have been acknowledged previously within the field, and there are also examples of excellent research (**Evidence that school leadership and environment matter**). Overall, however, there is so much that is poor that the field of school leadership research has some way to go before researchers can feel proud of its contribution and practitioners can trust its results.

References

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. In *Psychological Review* (Vol. 84, Issue 2, pp. 191–215). American Psychological Association. https://doi.org/10.1037/0033-295X.84.2.191

Barker, J., & Rees, T. (2020). What is school leadership? In S. Lock (Ed.), *The researchED guide to leadership:* An evidence informed guide for teachers. John Catt.

Bradley, G. W. (1978). Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of Personality and Social Psychology,* 36(1), 56–71. https://doi.org/10.1037/0022-3514.36.1.56

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organising schools for improvement: Lessons from Chicago*. The University of Chicago Press.

Ebel, R. L. (1951). Estimation of the reliability of ratings. Psychometrika, 16(4), 407–424. https://doi.org/10.1007/BF02288803

Elmore, R. F. (2000). Building a New Structure for School Leadership. In *Albert Shanker Institute.* Albert Shanker Institute. 555 New Jersey Avenue NW, Washington, DC 20001. Tel: 202-879-4401; Fax: 202-879-4403; Web site: http://www.shankerinstitute.org. https://eric.ed.gov/?id=ED546618

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron,* 1, 3–32.

Forer, B. R. (1949). The fallacy of personal validation: A classroom demonstration of gullibility. *The Journal of Abnormal and Social Psychology,* 44(1), 118–123. https://doi.org/10.1037/h0059240

Fullan, M. (2003). *The moral imperative of school leadership.* SAGE Publications. https://books.google.co.uk/books?id=LPf2Je8UDbcC

Gates, S., Baird, M., Doss, C. J., Hamilton, L., Opper, I. M., Master, B. K., Prado Tuma, A., Vuollo, M., & Zaber, M. A. (2019). *Preparing school leaders for success: Evaluation of New Leaders' Aspiring Principals program,* 2012-2017. RAND Corporation. https://doi.org/10.7249/RR2812

Gehlbach, H., & Artino, A. R. (2018). The survey checklist (Manifesto). *Academic Medicine, 93*(3), 360–366. https://doi.org/10.1097/ACM.0000000000002083

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology, 15*(4), 380–387. https://doi.org/10.1037/a0025704

Goldring, E., Huff, J., & Spillane, J. P. (2008). Measuring principals' content knowledge of learning-centered leadership. *Annual Conference of the American Educational Research Association.*

Grissom, J. A., Egalite, A. J., & Lindsay, C. A. (2021). *How principals affect students and schools: A systematic synthesis of two decades of research.* http://www.wallacefoundation.org/principalsynthesis.

Grissom, J. A., & Loeb, S. (2009). Triangulating principal effectiveness: How perspectives of parents, teachers, and assistant principals identify the central importance of managerial skills. In *National Center for Analysis of Longitudinal Data in Education Research* (No. 35). National Center for Analysis of Longitudinal Data in Education Research. The Urban Institute, 2100 M Street NW, Washington, DC 20037. Tel: 202-261-5739; Fax: 202-833-2477; e-mail: inquiry@caldercenter.org; Web site: http://www.caldercenter.org. www.caldercenter.org

Grissom, J. A., & Loeb, S. (2011). Triangulating principal effectiveness: How perspectives of parents, teachers, and assistant principals identify the central importance of managerial skills. *American Educational Research Journal,* 48(5), 1091–1123. https://doi.org/10.3102/0002831211402663

Hallinger, P. (2005). Instructional leadership and the school principal: A passing fancy that refuses to fade away. *Leadership and Policy in Schools, 4*(3), 221–239. https://doi.org/10.1080/15700760500244793

Hallinger, P., & Murphy, J. (1985). Assessing the instructional management behavior of principals. *The Elementary School Journal, 86*(2), 217–247. https://doi.org/10.1086/461445

Hallinger, P., & Wang, W.-C. (2015). *Assessing instructional leadership with the Principal Instructional Management Rating Scale.* Springer International Publishing. https://doi.org/10.1007/978-3-319-15533-3

Harris, A., Leithwood, K., Day, C., Sammons, P., & Hopkins, D. (2007). Distributed leadership and organizational change: Reviewing the evidence. *Journal of Educational Change, 8*(4), 337–347. https://doi.org/10.1007/s10833-007-9048-4

Herman, R., Gates, S. M., Arifkhanova, A., Barrett, M., Bega, A., Chavez-Herrerias, E. R., Han, E., Harris, M., Migacheva, K., Ross, R., Leschitz, J. T., & Wrabel, S. L. (2017). *School leadership interventions under the Every Student Succeeds Act: Evidence review: Updated and Expanded.* https://www.rand.org/content/dam/rand/pubs/research_reports/RR1500/RR1550-3/RAND_RR1550-3.pdf

Hitt, D. H., & Tucker, P. D. (2016). Systematic review of key leader practices found to influence student achievement. *Review of Educational Research, 86*(2), 531–569. https://doi.org/10.3102/0034654315614911

Jacob, R., Goddard, R., Kim, M., Miller, R., & Goddard, Y. (2015). Exploring the causal impact of the McREL Balanced Leadership Program on leadership, principal efficacy, instructional climate, educator turnover, and student achievement: *Educational Evaluation and Policy Analysis, 37*(3), 314–332. https://doi.org/10.3102/0162373714549620

Kelley, T. L. (1927). Interpretation of educational measurements. In *Interpretation of educational measurements.* World Book Co.

Kelvin, W. T. (1891). *Popular lectures and addresses* (Vol. 1). Macmillan.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality: *Education Researcher, 39*(8),

591–598. https://doi.org/10.3102/0013189X10390804

Leithwood, K. (2012). *The Ontario Leadership Framework 2012*. http://www.edu.gov.on.ca/eng/literacynumeracy/Framework_english.pdf

Leithwood, K., Harris, A., & Hopkins, D. (2019). Seven strong claims about successful school leadership revisited. *School Leadership & Management, 40*(1), 5–22. https://doi.org/10.1080/13632434.2019.1596077

Leithwood, K., & Jantzi, D. (2008). Linking leadership to student learning: *The contributions of leader efficacy. Educational Administration Quarterly,* 44(4), 496–528. https://doi.org/10.1177/0013161X08321501

Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *Review of research: How leadership influences student learning*. https://www.wallacefoundation.org/knowledge-center/Documents/How-Leadership-Influences-Student-Learning.pdf

Liebowitz, D. D., & Porter, L. (2019). The effect of principal behaviors on student, teacher, and school outcomes: A systematic review and meta-analysis of the empirical literature. *Review of Educational Research,* 89(5), *785–827*. https://doi.org/10.3102/0034654319866133

Lietz, P. (2010). Research into questionnaire design: A summary of the literature. *International Journal of Market Research, 52*(2), 249–272. https://doi.org/10.2501/S147078530920120X

Macey, E., Decker, J., & Eckes, S. (2009). The Knowledge is Power Program (KIPP): An analysis of one model's efforts to promote achievement in underserved communities. *Journal of School Choice, 3*(3), 212–241. https://doi.org/10.1080/15582150903280656

Marzano, R. J., Waters, T., & McNulty, B. A. (2005). *School leadership that works: From research to results*. Association for Supervision and Curriculum Development.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*(1), 28–50. https://doi.org/10.1177/1088868310366253

Meehl, P. E. (1956). Wanted—a good cook-book. *American Psychologist, 11*(6), 263–272. https://doi.org/10.1037/h0044164

Murphy, J. (2005). *Connecting teacher leadership and school improvement*. SAGE Publications. https://books.google.co.uk/books?id=Et3nOl5VVnIC

Pajares, F. (1997). Current directions in self-efficacy research. In *Advances in Motivation and Achievement, Volume 10*. JAI Press.

Protzko, J. (2022). Invariance: What does measurement invariance allow us to claim? *PsyArXiv*. https://doi.org/10.31234/OSF.IO/R8YKA

Robinson, V., & Gray, E. (2019). What difference does school leadership make to student outcomes? *Journal of the Royal Society of New Zealand, 49*(2), 171–187. https://doi.org/10.1080/03036758.2019.1582075

Robinson, V., Hohepa, M., & Lloyd, C. (2009). *School leadership and student outcomes: Identifying what works and why.* https://www.educationcounts.govt.nz/__data/assets/pdf_file/0015/60180/BES-Leadership-Web-updated-foreword-2015.pdf

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied, 80*(1), 1–28. https://doi.org/10.1037/h0092976

Thorndike, E. L. (1904). *An introduction to the theory of mental and societal measurements* (1st ed.). Science Press.

Tian, M., Risku, M., & Collin, K. (2016). A meta-analysis of distributed leadership from 2002 to 2013. *Educational Management Administration & Leadership, 44*(1), 146–164. https://doi.org/10.1177/1741143214558576

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*(2), 202–248. https://doi.org/10.3102/00346543068002202

Wang, H., & Hall, N. C. (2018). A systematic review of teachers' causal attributions: Prevalence, correlates, and consequences. *Frontiers in Psychology, 9.* https://doi.org/10.3389/fpsyg.2018.02305

Wilson, C. L. (2005). Principal leadership, school climate, and the distribution of leadership within the school community [University of Montana]. In *ProQuest Dissertations and Theses.* https://ezp.lib.cam.ac.uk/login?url=https://www.proquest.com/dissertations-theses/principal-leadership-school-climate-distribution/docview/305455978/se-2?accountid=9851

Witziers, B., Bosker, R. J., & Krüger, M. L. (2003). Educational leadership and student achievement: The elusive search for an association. *Educational Administration Quarterly, 39*(3), 398–425. https://doi.org/10.1177/0013161X03253411

Published by Evidence Based Education in partnership with Cambridge Assessment International Education