This is a repository copy of *Selecting teachers and prospective teachers:a meta-analysis*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/140347/

Version: Published Version

Article:
Klassen, Robert Mark orcid.org/0000-0002-1127-5777 and Kim, Lisa orcid.org/0000-0001-9724-2396 (2019) Selecting teachers and prospective teachers:a meta-analysis. Educational Research Review. pp. 32-51. ISSN 1747-938X

https://doi.org/10.1016/j.edurev.2018.12.003

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Educational Research Review 26 (2019) 32–51

Contents lists available at ScienceDirect

Educational Research Review

journal homepage: www.elsevier.com/locate/edurev

Review

Selecting teachers and prospective teachers: A meta-analysis

Robert M. Klassen*, Lisa E. Kim

University of York, United Kingdom

ARTICLE INFO

Keywords:
Teacher selection
Teacher characteristics
Personality
Initial teacher education
Recruitment

ABSTRACT

The purpose of this review article was to examine the methods used for the selection of teachers for employment and prospective teachers entering initial teacher education (ITE) programs, and to assess the predictive validity of these methods. We reviewed 32 studies reporting selection methods administered in high-stakes conditions and that included an external (not self-reported) teacher effectiveness outcome measure. The overall effect size was small but significant ($r = 0.12$, $p < .001$). Moderator analyses showed that academic and non-academic predictors were both significantly associated with teacher effectiveness measures and that effect sizes were small (but significant) for selection into employment and ITE. We conclude the review by proposing a research agenda that has the potential to enhance educational outcomes by improving the selection of prospective teachers.

1. Introduction

Does it matter how we select teachers for employment or how we select candidates for initial teacher education (ITE)? Much attention has been paid to improving the quality of teaching and of ITE in the United States (e.g., Atteberry, Loeb, & Wyckoff, 2015; Darling-Hammond, 2010; Jackson, Rockoff, & Staiger, 2014) and internationally (e.g., Hattie, 2009; Hobson, Ashby, McIntyre, & Malderez, 2010; OECD, 2005). Evaluating the effectiveness of practicing teachers (Goldhaber, 2015; Kane & Staiger, 2012) and the quality of ITE programs (Greenberg, McKee, & Walsh, 2015; U.S. Department of Education, 2016) have become national education priorities, but little systematic attention has been focused on how we make decisions about choosing teachers for the classroom or how we select prospective teachers for ITE programs. Furthermore, beliefs about teacher selection[1] may be influenced by a pervasive implicit belief that teaching is unique, and that lessons learned from selection research and methods in other disciplines do not apply to education (Harris & Rutledge, 2010).

Systematic research on selection into employment and training is well developed in a wide variety of fields: law (Marks & Moss, 2016), medicine (Patterson et al., 2016), the military (Gayton & Kehoe, 2015), business (Cook, 2009), undergraduate and graduate education (Richardson, Abraham, & Bond, 2012; Schwager, Hülsheger, Bridgeman, & Lang, 2015), and dentistry (Taylor, Mehra, Elley, Patterson, & Cousans, 2016). But systematic research on the selection of teachers and prospective teachers—a role weighted with enormous social and economic expectations and impact (Hanushek, 2014)—has largely been ignored, and selection methods in education have been described as "ad hoc" (p. 24, Goldhaber, Grout, & Huntington-Klein, 2014) and "information poor" (p. 324, Liu & Johnson, 2006). The purpose of this article is to review teacher and prospective teacher selection methods, to assess the predictive validity of these methods, and to propose how to improve future teacher and prospective teacher selection practices in ways that

* Corresponding author. Department of Education, University of York, Heslington, York, YO10 5DD, United Kingdom
E-mail address: robert.klassen@york.ac.uk (R.M. Klassen).

[1] For ease-of-use, we use the term "teacher selection" to refer to selection of teachers for employment and the selection of prospective teachers into ITE programs.

https://doi.org/10.1016/j.edurev.2018.12.003
Received 6 May 2018; Received in revised form 12 December 2018; Accepted 20 December 2018
Available online 26 December 2018
1747-938X/ © 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/BY/4.0/).

reflect the empirical and theoretical work conducted in education and other disciplines.

1.1. Selection methods

In an employment or training setting, selection methods are used in at least three situations: (a) when the number of applicants is greater than the number of available jobs or training places, (b) when there is a need to identify unsuitable applicants ('selecting out') before employment or training, and (c) when there is a desire to generate profiles of applicant strengths and weaknesses for future development. A selection process is built on an evaluation of available data; an important decision made by selectors is to determine the kinds of data that will be gathered in order to make predictions about future success. In the past, selection methods for most kinds of employment and training were left up to selector judgment and intuition, occasionally with the help of methods as diverse as astrology, physiognomy (analysis of facial features) and graphology (analysis of hand-writing; Chamorro-Premuzic & Furnham, 2010). Surprisingly, some of these unsubstantiated selection methods are still used for personnel selection (e.g., Bangerter, Konig, Blatti, & Salvisberg, 2009).

In education, those making selection decisions may be prone to a similar lack of critical analysis of the validity of current selection methods, with little systematic gathering of evidence for the efficacy of selection methods (Bowles, Hattie, Dinham, Scull, & Clinton, 2014; Liu & Johnson, 2006). In a recent study in the UK, Davies et al. (2016) interviewed those making selection decisions for ITE and for school-based teacher training programs, and found that selectors emphasized the importance of their intuition when making selection decisions: "Really, you do only need one question and a bit of conversation to find out what you are looking for" (p. 298), with selectors relying on a "gut feeling" to identify the "X factor" (p. 298). In spite of the apparent confidence in the interviewees' selection approaches, little evidence was offered for the quality of their selection decisions (e.g., "I wouldn't have any statistics … after they've left us," p. 297). Although most people are confident that they can accurately judge personal characteristics through interviews (Dana, Dawes, & Peterson, 2013), extensive research from organizational psychology suggests that interviewers are prone to unreliable judgment and are influenced in their decision-making by unconscious biases based on race, age, and appearance (Cook, 2009). Teacher selection models may be built on sophisticated teaching standards frameworks that include multiple domains of practice and competence (e.g., Casey & Childs, 2017; Denner, Salzman, & Newsome, 2001), but may fail to adopt selection methods that reliably assess these domains.

Selection methods can be expensive. Expenses for school districts or ITE programs may include personnel costs for selection activities (e.g., multiple personnel involved with conducting interviews, observing group activities, or scoring personal statements), fees for training those who administer commercial selection tools, and further costs associated with scoring and reports for selection instruments. Metzger and Wu (2008) argued that examining the validity of selection methods in education is important at least partly because of the costs accrued for each candidate. Selection practices require periodic reviews to ensure that the methods used are cost-effective, reliable (consistent over time), valid (show evidence of predictive utility), and fair (unbiased) for all applicants, regardless of gender, ethnicity, sexuality, and socio-economic status.

1.1.1. Selection research in other disciplines

Over the last 100 years, systematic research has explored the reliability, validity, and fairness of academic and non-academic selection methods in a range of disciplines (Gottheil & Michael, 1957; Ryan & Ployhart, 2014). A recent review of selection into medical training (Patterson et al., 2016) found that academic records, multiple mini-interviews (MMIs), and situational judgment tests (SJTs) were less prone to bias and more predictive of outcomes (e.g., medical school performance and supervisor ratings of trainee job performance) than reference letters and individual interviews. The predictive validity of selection methods varies across disciplines: predictive validity coefficients (i.e., correlations with outcomes such as supervisor ratings of performance) range from 0.18 to 0.43 in dentistry (Patterson, Ashworth, Mehra, & Falcon, 2012), 0.37 in medicine (Lievens & Patterson, 2011), between 0.06 and 0.50 in business (Christian, Edwards, & Bradley, 2010), and 0.34 across multiple occupation groups (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Critical analysis of selection methods in other disciplines is growing (e.g., clinical psychology; Woolf et al., 2015), but this research growth has not been seen in education.

1.1.2. Teacher selection research and practices

The lack of systematic research on selection into teaching—perhaps unique among major professional fields—is surprising in view of recent efforts on identifying and measuring the key factors of teacher effectiveness (e.g., Kane, McCaffrey, Miller, & Staiger, 2013), tracing the effectiveness trajectories of new teachers (e.g., Atteberry et al., 2015), and on identifying teacher characteristics linked with teaching outcomes (e.g., Robertson-Kraft & Duckworth, 2014). Although pockets of research on selection of teachers for ITE and employment have been conducted by education researchers over the past decades (e.g., Ebmeier & Ng, 2005; Metzger & Wu, 2008; Wise, Darling-Hammond, & Berry, 1987), systematic research has been sporadic. A 2008 meta-analysis explored the predictive validity of one widely used commercial selection tool, the Teacher Perceiver Interview (TPI; Metzger & Wu, 2008). The authors analyzed 24 studies (16 dissertations, 7 reports from the test publisher, and 1 journal article) and found a range of -0.12 to 0.87 for the correlations between TPI scores and indicators of teaching effectiveness, with a weighted mean of $r = 0.28$, and a median r of 0.22, considered by the authors to be a 'modest' effect size. In order to understand the antecedents of teacher effectiveness, it is helpful to consider the value and validity of current selection practices, and to explore viable alternatives used in other disciplines.

Due to a lack of systematic research on selection in education, the methods in place often lack supporting evidence for their effectiveness. A recent review of teacher education selection practices in England and Wales (Klassen & Dolan, 2015) showed that most programs use the same predictors—personal statements, letters of reference, and interviews—to evaluate the personal

characteristics of candidates. But evidence from other professional fields, especially medicine, suggests that evidence for personal statements is "mixed at best," (Patterson et al., 2016, p. 42), that letters of reference "are neither a reliable nor a valid tool" for selection (p. 43), and that traditional face-to-face interviews are unreliable and prone to a range of systematic biases (Kausel, Culbertson, & Madrid, 2016).

1.2. Predicting teacher effectiveness

Making decisions about selecting prospective teachers is, at its heart, making a prediction about future teacher effectiveness. The selection process into employment and into ITE addresses the same fundamental predictive question: *Is it likely that this applicant will be (or develop into) an effective teacher?* Teachers become more reflective, more knowledgeable about teaching, and generally more effective as they gain experience (e.g., Antoniou, Kyriakides, & Creemers, 2015; Atteberry et al., 2015; Bahr & Mellor, 2016). Although some characteristics for selection into employment and into ITE may differ in kind and degree (e.g., employers expect applicants to have specific pedagogical knowledge that ITE selectors would not expect in applicants), other characteristics (e.g., reasoning ability, motivation, empathy, conscientiousness) are similarly important across the teaching experience continuum (Kunter, Kleickmann, Klusmann, & Richter, 2013).

Teacher effectiveness—during ITE or during professional practice—is difficult to define and even more difficult to measure (e.g., Kane et al., 2013). Researchers and practitioners have used observation measures, student value-added scores, and student reports to assess teacher effectiveness, with sometimes limited agreement between the different forms of assessment (Chetty, Friedman, & Rockoff, 2014; Grissom & Loeb, 2017). Recent systematic research on measuring teacher effectiveness has provided support for measures such as the Classroom Assessment Scoring System (CLASS; Hamre et al., 2013), the International Comparative Analysis of Learning and Teaching (ICALT; van der Lans, van de Grift, & van Veen, 2018), and the International System for Teacher Observation and Feedback (ISTOF; Muijs et al., 2018). Research on teacher effectiveness measures has increased in the last decade, with improved reliability and validity, and increased consensus on the factors that underpin the frameworks and models of effective instructional practice (Gill, Shoji, Coen, & Place, 2016; Klassen & Tze, 2014).

1.2.1. Models of teacher effectiveness

Teacher effectiveness refers to a set of within-person attributes—personality, motivation, beliefs, and dispositions—that interact with contextual factors (cultural, social, educational) to influence student outcomes. Some models of teacher effectiveness focus largely on the role of learning opportunities, including teacher education and professional learning (e.g., Muijs et al., 2014). Other models focus largely on within-classroom behaviors, such as the *Teaching through Interactions* model proposed by Hamre et al. (2013). Kunter and colleagues' COACTIV model (see Fig. 1; Kunter et al., 2013) provides a useful framework for teacher selection because it is built on a dynamic interactionist view (Steyer, Schmitt, & Eid, 1999) that suggests that the expression of traits and attributes depends on the interface between within-person and contextual factors. The COACTIV model proposes that teaching competence develops over time through exposure to learning opportunities, but it also acknowledges the influence of within-person factors that are present at entry into teacher training and practice. The active engagement in, and reflections on, learning opportunities are not just dependent on the quality of the opportunities, but are also influenced by the characteristics of the individual to whom the opportunity is presented.

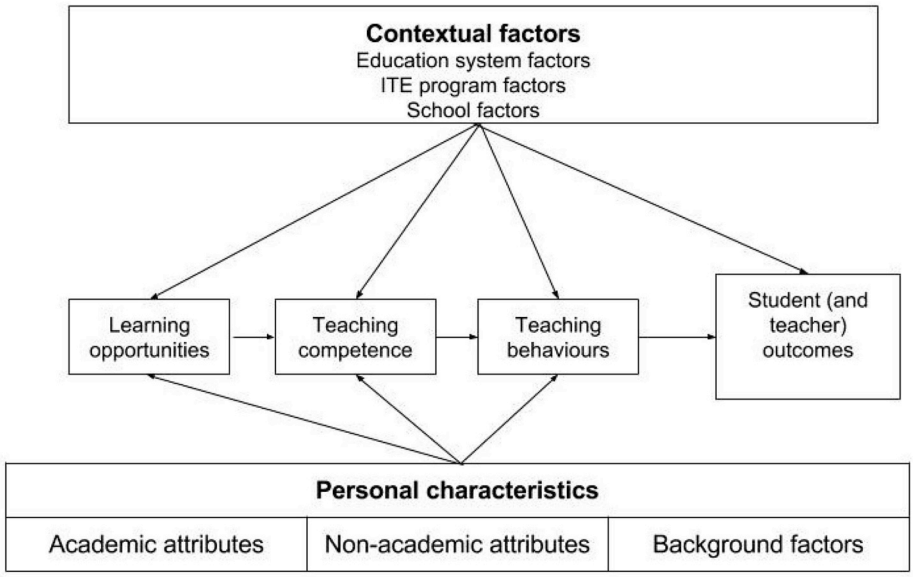

Fig. 1. Dynamic interactionist view of teacher effectiveness (adapted from Kunter et al., 2013).

Key within-person factors include academic attributes (in some literature called 'cognitive attributes') such as subject area knowledge, reasoning abilities, and pedagogical knowledge. Academic attributes can be readily assessed using existing academic records, subject-related tests, or various well-researched measures of cognitive abilities. Non-academic attributes (sometimes referred to as 'non-cognitive' attributes) refer to beliefs, motives, personality traits, and dispositions (e.g., Patterson et al., 2016). In high-stakes testing, such as for selection purposes, measuring non-academic attributes presents unique challenges due to the prevalence of socially desirable responding and faking (Sackett, Lievens, Van Iddekinge, & Kuncel, 2017). Background factors such as educational qualifications and teaching-related experiences also form part of the personal characteristics that directly and indirectly influence teacher effectiveness. Kunter et al., 's 2013 model provides a theoretical foundation to explain how contextual factors, learning opportunities, and personal characteristics are all relevant to decisions about selecting prospective teachers.

1.3. The importance of research on teacher selection

Selecting prospective teachers for employment and teacher education programs can lead to improvements in educational outcomes because teacher effectiveness is related to student achievement, and because effectiveness varies among teachers (Darling-Hammond & Youngs, 2002; Hindman & Stronge, 2009). Furthermore, teacher effects on student learning are additive and cumulative over time, meaning that a sequence of less effective teachers is difficult for a student to overcome (Wright, Horn, & Sanders, 1997). The selection of teachers and teacher candidates is important not just for students' academic outcomes and well-being, but for a nation's social and economic well-being. Economists rate the impact of a single teacher on class-level lifetime earnings gains in the hundreds of thousands of dollars (e.g., Hanushek & Rivkin, 2012; Kane & Staiger, 2012), and hiring a new teacher represents a career-long investment of at least two million dollars (Goldhaber et al., 2014), meaning that selecting a teacher who is less effective, or who leaves the system (Synar & Maiden, 2012), may represent a costly mistake.

Although almost all new teachers and teacher education candidates will become more effective as they gain experience (e.g., Harris & Rutledge, 2010; Staiger & Rockoff, 2010), it is also true that the 'effectiveness starting point' may be important, with the gap between higher- and lower-performing teachers showing stability over time. Atteberry et al. (2015) found that the initial performance of new teachers was reliably predictive of later performance, suggesting that early identification, i.e., selection, of the most effective teachers pays long-term dividends. Some candidates will show higher levels of growth and teaching effectiveness than others, and some candidates in ITE programs experience little change in teaching-related motivation over time (Klassen & Durksen, 2014). Watt, Richardson, and Wilkins (2014) traced the motivation of students from the beginning to the end of their ITE program, and found that a sizable group—28% of participants in their study—began the program with low motivation and maintained that negative profile throughout the program. Predictable variation in teacher effectiveness (Atteberry et al.; Weisberg, Sexton, Mulhern, & Keeling, 2009) suggests that selection decisions have long-term consequences and that examining the effectiveness of selection methods is worthwhile.

Countries with high-performing education systems, like Singapore (Sclafani, 2015) and Finland (Sahlberg, 2015), tend to place high levels of importance on developing rigorous approaches for teacher and prospective teacher selection (Barber & Mourshed, 2007; Hobson et al., 2010). In education systems with less heated competition for ITE places, there have also been calls for attention to improving selection methods, (Heinz, 2013; Thomson et al., 2011; UK House of Commons, 2012). Selection is necessary when the number of applicants outweighs the number of available places, but is also helpful to identify those candidates who may be unsuitable for a particular role (i.e., 'selecting out'), and to identify candidates' strengths and weaknesses for future development.

1.4. Models of teacher selection

The predictors used for teacher selection can be chosen because they are hypothesized to target one or more target outcomes. Fig. 2 provides a model (adapted from Ployhart & Schneider, 2012) that shows how selection methods are built on *inferred* relations. In this model, the latent constructs are represented by circles, and the observed variables (the measures used in selection) are represented by rectangles, with the inferred relations represented by dashed arrows. Arrow 4 represents the empirical relation between a selection measure (e.g., a measure of personality or motivation) and an outcome measure (e.g., a score from teaching observations), usually expressed as a correlation between the selection measure (i.e., predictor) and scores on a teaching outcome measure (i.e., criterion). Arrow 2 mirrors the relation between the predictor and outcome but denotes its theoretical or latent relation. Arrows 1 and 3 represent construct validity: arrow 1 represents the extent to which the selection measure (predictor) represents the construct of interest, and arrow 3 represents the construct validity of the outcome measure, and whether the outcome, usually some kind of measure of teacher effectiveness, represents the person's 'true' teaching effectiveness. Arrow 5 lies at the heart of the selection process and indicates the degree to which scores from an imperfect selection measure (e.g., measure of motivation or personality, letter of reference, face-to-face interview) are associated with 'true' differences in teacher effectiveness, imperfectly measured by an outcome measure (e.g., grade on teaching practicum). Although the relation between the actual measure used in selection and the latent teacher effectiveness variable cannot be directly assessed, it can be inferred through the other relations (arrows) described in the model.

1.5. The present study

This review focuses on methods of selection of teachers (into employment) and prospective teachers (into ITE programs). The first goal of the review is to systematically examine recent research (i.e., since 2000) exploring selection methods for teaching and

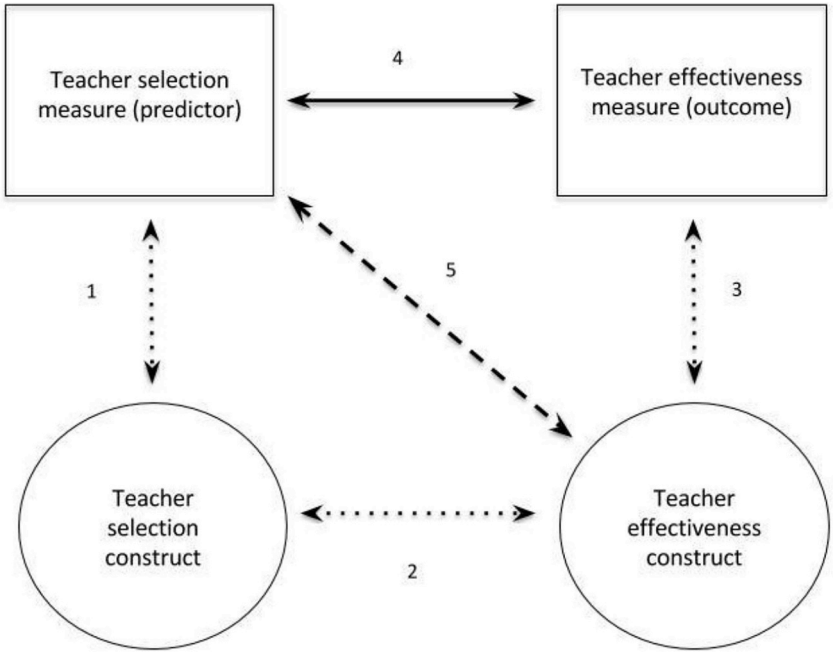

Fig. 2. Teacher selection model (adapted from Ployhart & Schneider, 2012).

selection methods for teacher education. In particular, we are interested in the predictive validity of the methods used for teacher selection. The second goal is to use the information gained from the review and to combine it with recent advances in selection methods outside of education in order to consider improved selection methods in education.

The following questions are posed:

1. What is the predictive validity of the methods used to select teachers and teacher candidates (i.e., what is the statistical relationship between selection methods and measured teacher effectiveness)?
   (a) Are there differences in the predictive validity of tests assessing academic and non-academic constructs?
   (b) Are there differences in the prediction of different kinds of teacher effectiveness measures, i.e., teacher observation measures and student performance measures, and practicum reports?
   (c) Are there differences in the predictive validity of the methods used for selection into employment and for selection for ITE programs?
2. What is the relation between cost and benefit (i.e., predictive validity) of selection practices?

2. Method

2.1. Literature search

Search procedures for systematic reviews should be reproducible, transparent, and thorough, and should identify relevant studies from a wide range of sources (Booth, Sutton, & Papaioannou, 2016). We conducted an exhaustive search of the literature using five search strategies. First, we examined the contents of previous literature reviews of teacher selection (e.g., Ebmeier, Dillon, & Ng, 2007; Goldhaber et al., 2014; Metzger & Wu, 2008) for articles that fit our search criteria (i.e., that focused on *teacher selection* and *selection/admissions into teacher education/teacher training*). Second, we searched a range of databases (British Education Index, ERIC, PsycINFO, Google Scholar, ProQuest Dissertations and Theses Global, and EconLit) with combinations of *teacher selection, teacher hiring, teacher education + admissions, teacher hiring instruments*. Third, we hand-searched five high-ranking education-related review journals (*Educational Psychology Review, Educational Research Review, Review of Educational Research,* and *Educational Psychologist*) for relevant studies. Fourth, we hand-searched the reference list and citations of each of the articles that met our inclusion criteria. Fifth, we emailed authors of relevant studies requesting further data where bivariate statistics between predictor and criterion variables were not calculable (e.g., in multiple regression using more than one predictor variable).

2.2. Inclusion and exclusion criteria

The focus of our review was on selection methods used during actual selection processes, not during exploratory research studies, low-stakes pilot studies, or during development of selection tools. How a participant responds to an assessment exercise depends on

the consequences of the experience: participants respond differently to selection tools administered in high-stakes and low-stakes settings (Abdelfattah, 2010). We included studies that:

- Reported primary research in the form of journal articles, dissertations, and published research reports (including publicly available working papers, such as National Bureau of Economic Research working papers)
- Were published between 2000 and 2017. We chose 2000 as a starting date because Metzger and Wu's (2008) meta-analysis largely focused on pre-2000 studies, with only two peer-reviewed publications post-2000
- Included participants who were candidates for K-12 teaching positions or for ITE programs
- Included an academic or non-academic measure assessed at the point of selection for teaching employment or for admission into an ITE program
- Included an outcome measure that measured teacher effectiveness using an external measurement (i.e., not self-reported). Examples of outcome measures are teaching observation ratings (e.g., a principal's rating of a practicing teacher's performance, or a supervisor's rating of a pre-service teacher's teaching during a practicum), and classroom-level student achievement gain (e.g., value-added) scores. In ITE settings we included scores or grades for teaching practice given by university supervisors or teaching mentors, but not for overall academic performance (i.e., overall GPA in a teacher education program).

We excluded studies that:

- Examined the relation between *practicing* teachers' performance on selection methods, i.e., administered for research and development—not selection— purposes (e.g., Ebmeier & Ng, 2005)
- Reported evaluations of teacher selection tools without reporting data gathered from administration of said tools (e.g., Mertz, 2010)
- Reported unverified data provided by test companies that provided non-peer-reviewed validity data supporting their proprietary selection methods (e.g., Gallup Organization, 2001).

2.3. Search results

Our initial literature search (from databases and hand-searches) resulted in 1306 records. As a first screen, we deleted duplicates, and then scanned titles and abstracts for all of the remaining records for relevance to our study questions, resulting in a pool of 498 articles to analyze more closely. The second screen applied the inclusion and exclusion criteria to each of the studies, resulting in a pool of 51 studies. For the third screen, we downloaded and carefully read the 51 studies (journal articles, dissertation reports, and published working papers) to see if a bivariate statistic could be computed between selection measure and outcome measure. Records were excluded from further analysis if bivariate statistics (e.g., Pearson's r, Cohen's d, Hedge's g, Glass' Δ, multiple regression with only one IV) were not reported or could not be calculated, as per procedures of other meta-analyses (e.g., Kuykendall, Tay, & Ng, 2015). Meta-analyses necessarily focus on bivariate relations and cannot aggregate partial correlation coefficients (as in multiple regression) unless all studies use the same set of IVs. However, the role of some intervening variables can be calculated, as we did here, through moderator analysis.

After applying the three levels of filtering to our initial record of 1306 records, we identified 32 studies that met the criteria for inclusion in our review. The studies in our review are largely distinct from the 24 studies included in the Metzger and Wu review of commercial teacher selection instruments (2008), with only two studies (Buresh, 2003; Young & Delli, 2002) overlapping. We decided to retain the two overlapping studies in order to fully capture the range of articles for our target timeframe, i.e., 2000–2017. The flow chart in Fig. 3 provides a visual overview of the search process.

2.4. Coding

A coding procedure was developed to analyze the included studies. The two authors jointly developed a coding protocol to capture the key study criteria. They independently coded a sample of the studies with a high level of agreement (nearly 100%, reflecting the objective nature of the coding criteria). A consensus discussion approach was used to resolve any differences about interpretation of coding of predictor or outcome variables. Correlations were reverse scored if necessary (e.g., when an outcome measure included a scale with ‘1’ as a positive score and ‘5’ as a negative score), so that higher scores represented higher levels of the teacher effectiveness outcome measure. We coded the 32 studies that met our search criteria for:

1. Reference details (i.e., dissertation, journal article, published report)
2. Study location (country of data collection)
3. Selection level (selection for employment, or selection for ITE, either primary, secondary, or both)
4. Selection method used (i.e., predictor variable)
5. Classification of predictor variable (tests assessing academic, non-academic, or both types of constructs)
6. Classification of outcome measure of teacher effectiveness used in study (e.g., ratings derived from principal observation, student value-added learning scores)
7. Results (qualitative and quantitative [effect size] description of predictive validity)
8. Estimation of cost of implementation (time cost, training and administration costs) for administering selection method.

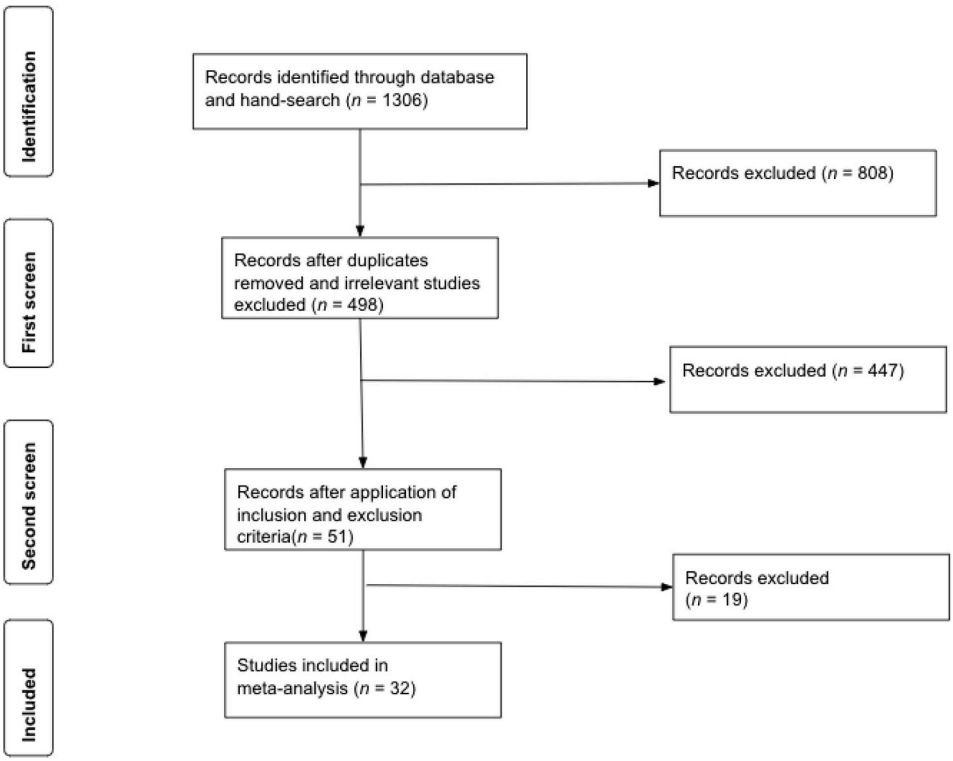

Fig. 3. Flow chart of the inclusion and exclusion of studies.

2.4.1. Estimation of the cost of implementation

For the estimation of cost of implementation of selection method, we calculated the average cost necessary to assess each candidate using the selection method described in the study. We summed three components to calculate the estimated average cost per candidate: (a) an assessor's initial training costs divided by the number of candidates that an assessor was estimated to evaluate (e.g., 30 candidates), (b) estimated cost for the time spent for the assessor to be trained, and (c) estimated cost for the time spent evaluating the candidates. The cost for the training for commercial training tools was calculated from publicly available information, which took into account the length of training periods (e.g., Teacher Perceiver Interview, Haberman Star Interview). The costs for some methods could not be calculated as the studies did not specify the length of training (e.g., Venture for Excellence).

When details of the assessment time length were not specified, estimates were applied: group discussions were estimated to take 45 min, structured interviews were estimated to take 30 min, checking applications (e.g., resumes, letters of recommendation), and checking for academic results were estimated to take 15 min for each task. Hourly rates based on the latest statistics available from OECD's teachers' statutory salaries divided by the statutory teaching hours per school year in public institutions for upper secondary general education at the top of salary scale (OECD, 2017) were applied. In effect, a pay rate of $69.89/hour was applied to the hourly cost estimations for the US-based studies, $99.77 for Switzerland, $88.32 for Canada, and $88.35 for Ireland (all in US dollars). If multiple assessors were necessary for assessing a candidate, the cost for multiple assessors was calculated. If multiple candidates were assessed simultaneously (e.g., in a group activity), the cost of assessing each candidate (that is, dividing the cost of having the assessor (s) by the number of candidates present) was calculated. Training time was included when stated in the methods, whereby a day of training was estimated to take 6 h per day. Information on costs could not be found for McWhirter's (2014) and Oestreich's (2016) studies and hence these studies were not included in the cost analysis.

2.4.2. Statistical correction

The Meta-Essentials program (van Rhee, Suurmond, & Hak, 2015) was used to calculate effect sizes in the meta-analysis. We used an inverse variance-weighting random effects model which assumes that 'true' effect sizes vary across studies. To weight multiple correlations within each study equally across multiple studies, individual Pearson *r*s were transformed into Fisher's *z* scores, which were then averaged within each study, after which the average Fisher's *z* scores were transformed back to Pearson *r*s (Pearson *r*s were converted to *z*-scores because *r*s with small sample sizes are less likely to be normally distributed). These transformed Pearson *r*s were then used to calculate the final effect sizes for the meta-analysis.

2.4.3. Interpreting effect sizes

Hattie suggests the following effect size descriptions for educational researchd = 0.20 (roughly equivalent to r = 0.10) describes a small effect, d = 0.40 (r = 0.20) describes a medium effect, and d = 0.60 (r = 0.30) describes a large effect. However,

Table 1
Study descriptions and outcomes.

Authors and date	Source	Selection level	Sample size and location	Predictors (selection methods used)	Type of predictor	Outcome measures	Predictor - outcomes correlations	Est. cost per applicant
Bieri and Schuler (2011)	Assessment and Evaluation in Higher Education	ITE: elementary and secondary	*N* = 147 (Switzerland)	Assessment center assessing five competencies (communication, cooperation, assertiveness, motivation, fact finding) and self-reports, measured through a range of well-developed activities (including interviews, group task, and a presentation)	Non-acad	Self-report and mentor scores on a questionnaire based on the five assessment center dimensions; plus a mentor question: is this student a 'suitable teacher'? Plus test results in math, language, education and psychology	0.22	$289.61
Bingham (2000)	Dissertation	Teaching: elementary	*N* = 138 (U.S.A.)	Teacher Perceiver Interview (Gallup)	Non-acad	TAAS reading and math tests	0.03	$209.13
Buresh (2003)	Dissertation	Teaching: elementary	*N* = 851 (U.S.A.)	Teacher Perceiver Interview (Gallup)	Non-acad	Teacher evaluation scored by three experts who evaluated principal reports of summative teacher evaluations. 0–10 scale, plus absenteeism.	0.20	$209.13
Byrnes, Kiger, and Shechtman (2003)	Journal of Teacher Education	ITE: elementary	*N* = 68 (U.S.A.)	1. Academic performance before selection. 2. Group assessment procedure to measure verbal abilities, interpersonal skills, and leadership qualities and overall impression. Rated by 2 trained assessors	Both	Std teacher evaluation scores given by cooperating teacher and university supervisor measuring eight criteria. Inter-rater reliability was about .5–.6.	0.09	$26.21
Casebolt (2016)	Dissertation	Teaching: elementary and secondary	*N* = 271 (U.S.A.)	Teacher Selection Model online questionnaire	Academic	TEAM effectiveness and observation scores TVAAS value-added score	0.14 for student value-added 0.10 for observation	$69.89
Casey and Childs (2011)	Canadian Journal of Education	ITE: secondary/ math	*N* = 102 (Canada)	Undergrad GPA and written profile (e.g., 'describe three significant teaching-related experiences' and 'How might you contribute, etc.' scored by two reviewers on a 9-point scale	Both	Post hoc scoring of reports of practice teaching on a 13-pt scale and 5 subscales of judgments of 'preparedness for teaching' by three raters, including student	0.30	$88.32
Caskey, Peterson, and Temple (2001)	Teacher Education Quarterly	ITE	*N* = 141 (U.S.A.)	Five selection measures: GPA, letters of recommendation to test interpersonal skills, personal statement, standardized test scores for basic skills, overall file rating, performance on a group problem-solving activity	Both	Scoring at end of program including instructor reports, supervisor reports, grades, and teaching observations. Participants were rated in quintiles.	0.20	$97.84
Chesley (2004)	Dissertation	Teaching: elementary and secondary	*N* = 578 (U.S.A.)	Urban TPI screener version (i.e., 2/5 questions from 11 themes) and Praxis 1 score	Both	Teacher Performance Assessment system includes four teaching dimensions (e.g., 'designing knowledge work')	0.12	$164.87
Clemons (2010)	Dissertation	Teaching: elementary and secondary	*N* = 79 (U.S.A.)	Ventures for Excellence interview tool: 22-question interview with 11 categories (2 points per response)	Non-acad	Summative evaluations based on categories from the VFE tool and included teaching techniques, classroom management, interpersonal relationships, and professional responsibilities rated on a 1–5 scale	0.18	$131.18
Denner et al. (2009)	Educ Asse Eval Acc	ITE			Academic		0.24	0

(*continued on next page*)

Table 1 (*continued*)

Authors and date	Source	Selection level	Sample size and location	Predictors (selection methods used)	Type of predictor	Outcome measures	Predictor - outcomes correlations	Est. cost per applicant
			N = 452 (U.S.A.)	Praxis 1 scores in reading, writing, math (for entry into ITE), and admissions GPA		Renaissance Teacher Work Sample (TWS) scores which is evaluation of teaching with 8 standards in the 3rd and 4th year of ITE program		
Evans (2016)	Dissertation	Teaching: elementary and middle	*N* = 76 (U.S.A.)	Gallup TeacherInsight	Non-acad	Student (TGA) gain in math and reading	0.11 for student gain scores in reading; 0.10 for math	$53.07
Fry (2013)	Dissertation	Teaching: elementary and secondary	*N* = 452 (U.S.A.)	Gallup Teacher Insight (scored from 1 to 99)	Non-acad	Teacher Growth Index, a value-added score	0.08	$53.07
Gimbert and Chesley (2009)	Journal of School Leadership	Teaching: elementary and secondary	*N* = 578 (U.S.A.)	UTPI Urban teacher perceiver interview screener and Academic Skills Assesssment from Praxis	Both	(TPA) Teacher assessment scores based on four domains (e.g., creating a safe and orderly environment)	0.12	$164.87
Heinz (2013)	Educ Asse Eval Acc	ITE	*N* = 1023 (Ireland)	Primary degree result and prior teaching experience	Academic	Academic and teaching practice grades at end of PDE course	0.10	0
Henry et al. (2013)	Journal of Teacher Education	ITE	*N* = 279 (USA)	Praxis 1 scores	Non-acad	Math and reading value-added	0.01	0
Juracek (2017)	Dissertation	Teaching: elementary	*N* = 56 (U.S.A.)	Ventures Teacher Style Profile Builder (Form A Interview)	Non-acad	Student reading achievement	0.11 in reading and 0.10 in math	$131.18
Koerner (2007)	Dissertation	Teaching: elementary and secondary	*N* = 132 (U.S.A.)	TeacherInsight scores (Gallup), plus gender, age, degree). This TI is focused on reducing staff time for interviews, and can be done electronically. Requires 40 min to complete online using Likert scales. and open-ended questions	Non-acad	Texas Growth Index–an estimate of students' academic growth in reading, English and math (value-added)	0.21	$53.07
Kosnik, Brown, and Beck (2005)	The New Educator	ITE: elementary and secondary	*N* = 347 (Canada)	GPA + written profile about 3 sustained teaching experiences which are graded by two raters	Both	Practicum success and academic success, both derived from ratings by a supervisor/coordinator. Each of the two outcomes consisted of a set of criteria (e.g., classroom management) scored on a 1–4 scale.	0.14	$34.94
Kreifels (2003)	Dissertation	Teaching: elementary	*N* = 17 (U.S.A.)	Gallup TPI three 'behavioral subscales': intrapersonal, interpersonal, extrapersonal	Non-acad	Reading and math scores from state assessments plus Iowa test of basic skills	0.06	$125.29
Loyd (2004)	Dissertation	Teaching: elementary and middle school	*N* = 242 (U.S.A.)	Urban TPI	Non-acad	State scores in math, reading, or language	0.05	$203.30
McKinley (2009)	Dissertation	Teaching	*N* = 107 (U.S.A.)	Score on 7 written application questions, score on 32 question online StyleProfile, and score on 22-question interview	Non-acad	Scores on a rating scale for 1st-year teachers completed by school principals. The outcome measure was based on the 22-question interview	0.20	$131.18
McWhirter (2014)	Dissertation	Teaching: elementary and secondary	*N* = 50 (U.S.A.)	Teacher Selection Interview (Crown Global Consulting)	Non-acad	Tennessee Education Acceleration Model (TEAM) rubric administered by principals, based on Charlotte Danielson work, and TVAAS change in std achievement test scores	0.07	Not available

(*continued on next page*)

Table 1 (*continued*)

Authors and date	Source	Selection level	Sample size and location	Predictors (selection methods used)	Type of predictor	Outcome measures	Predictor - outcomes correlations	Est. cost per applicant
Mikitovics and Crehan (2002)	Journal of Educational Research	ITE: elementary and secondary	*N* = 1434 (U.S.A.)	Pre-Professional Skills Test (PPST) (which I think is same as Praxis 1) plus ACT scores. The PPST is composed of 3 subtests assessing reading, writing and math.	Academic	GPA and student teaching rating	0.15	0
Mills (2007)	Dissertation	Teaching: elementary	*N* = 116 (U.S.A.)	Ventures for excellence: 22 questions with 11 themes in 3 domains: purpose, human interaction, and teaching/ learning.	Non-acad	MAP tests which measure 'class growth' in language, reading, math	0.32	$131.18
Nelson (2013)	Dissertation	Teaching: elementary and secondary	*N* = 90 (U.S.A.)	Ventures for Excellence Teacher Style Profile: a 32 question web-based test measuring 13 themes: purpose, communicative, personable, compassionate, etc.	Non-acad	1. Teacher evaluation data in year one measuring instructional process, interpersonal relationships, and professional responsibilities on a 3-point scale. 2. Probationary teacher feedback scale scored by principals for purpose, human interaction, teaching/ learning, and overall effectiveness.	−0.05	$53.07
Novotny (2009)	Dissertation	Teaching: elementary and secondary	*N* = 527 (U.S.A.)	TeacherInsight scores. Web-based assessment multiple choice and Likert items requiring 30 min to complete. Produces a score from 0 to 100. Range was 43–89, with a mean of 66.4.	Non-acad	PDAS completed by 'an administrator' with scores in 8 domains: e.g., learner-centered instruction, active student participation, professional communication.	0.07	$53.07
Oestreich (2016)	Dissertation	Teaching: elementary and secondary	*N* = 88 (U.S.A.)	Applitrack TeacherFit screening assessment tool	Non-acad	Teacher Performance Evaluation observation tool	0.14	Not available
Robison (2003)	Dissertation	Teaching: elementary	*N* = 14 (U.S.A.)	Ventures for Excellence	Non-acad	Std growth in language, math, reading measured by NWEA RIT growth score	−0.32	$131.18
Ross (2013)	Dissertation	Teaching: elementary and secondary	*N* = 37 (U.S.A.)	Ventures for Excellence interview tool: 22-question interview with 11 categories (2 points per response)	Non-acad	Std achievement using the Ohio achievement test scores in math and reading	−0.17	$131.18
Sawyer (2005)	Dissertation	Teaching	*N* = 33 (U.S.A.)	Haberman Star Teacher Interview	Non-acad	Principal rating during first year using seven categories	−0.05	$131.53
Winters (2017)	Dissertation	Teaching: elementary and middle	*N* = 69 (U.S.A.)	Ventures for Excellence B-22	Non-acad	Staff Evaluation Instrument (based on observation)	0.11	$53.07
Young and Delli (2002)	Educational Administration Quarterly	Teaching	*N* = 196 (U.S.A.)	Teacher Perceiver Interview (Gallup)	Non-acad	Principal ratings of teachers using TPI domains (e.g., empathy, listening, etc.) using a rating system designed for this study, plus work attendance/ absenteeism	0.04	$209.13

Note. 'std' = 'student'; 'non-acad' = 'non-academic'.

interpretations of the practical value of effect sizes in education are fluid: Valentine and colleagues (Valentine, DuBois, & Cooper, 2004) considered an effect size of r = 0.08 between students' self-beliefs and subsequent achievement as a "small but noteworthy finding" (p. 127), and Coe (2002) proposed that an effect size of d = 0.10 (roughly r = 0.05) can result in important educational outcomes if the effect can be applies to all students (i.e., as in an effect involving teachers) and is cumulative over time.

3. Results

3.1. Description of studies

A summary of all of the included studies (reporting authors and date, education level, sample size and location, selection methods used, type of predictor, outcome measure used, predictor-outcome correlation, and cost per applicant for selection) is reported in Table 1. The majority of the studies (66% or 21/32) were PhD dissertations, with 34% (11/32) published in academic journals. Most of the studies—88%—were conducted in the United States (28/32), with the remaining studies conducted in Canada (2), Ireland (1), and Switzerland (1). The majority of the studies (72% or 23/32) investigated the relation between selection practices at the point of hiring and subsequent teaching effectiveness, and 28% (9/32) of the studies examined the relation between selection into teacher education and teaching effectiveness during the ITE program. Sample sizes ranged from 14 to 1434, $M = 274.69$, $SD = 326.01$.

We divided selection methods into three categories: *academic* (e.g., measures of academic achievement such as secondary school or university GPA, scores on academic tests such as Praxis I Math tests), *non-academic* (measures of beliefs, attitudes, personality, or motivation such as Gallup Teacher Perceiver Interview), and *both* (selection methods including both academic and non-academic measures). Most of the studies reported non-academic measures (23 or 72%), followed by both academic and non-academic (6 or 19%), and academic measures only (3 studies, or 9%). Selection measures included a range of commercial selection tests (e.g., Gallup TeacherInsight, Gallup Teacher Perceiver Interview, Ventures for Excellence, Teacher Selection Interview) and non-commercial methods (e.g., assessment center methods, letters of recommendation, personal statements).

3.2. Effect sizes

There were 32 studies included in the meta-analysis with an overall effect size of $r = 0.12$ ($df = 31$, $p < .001$), and a 95%

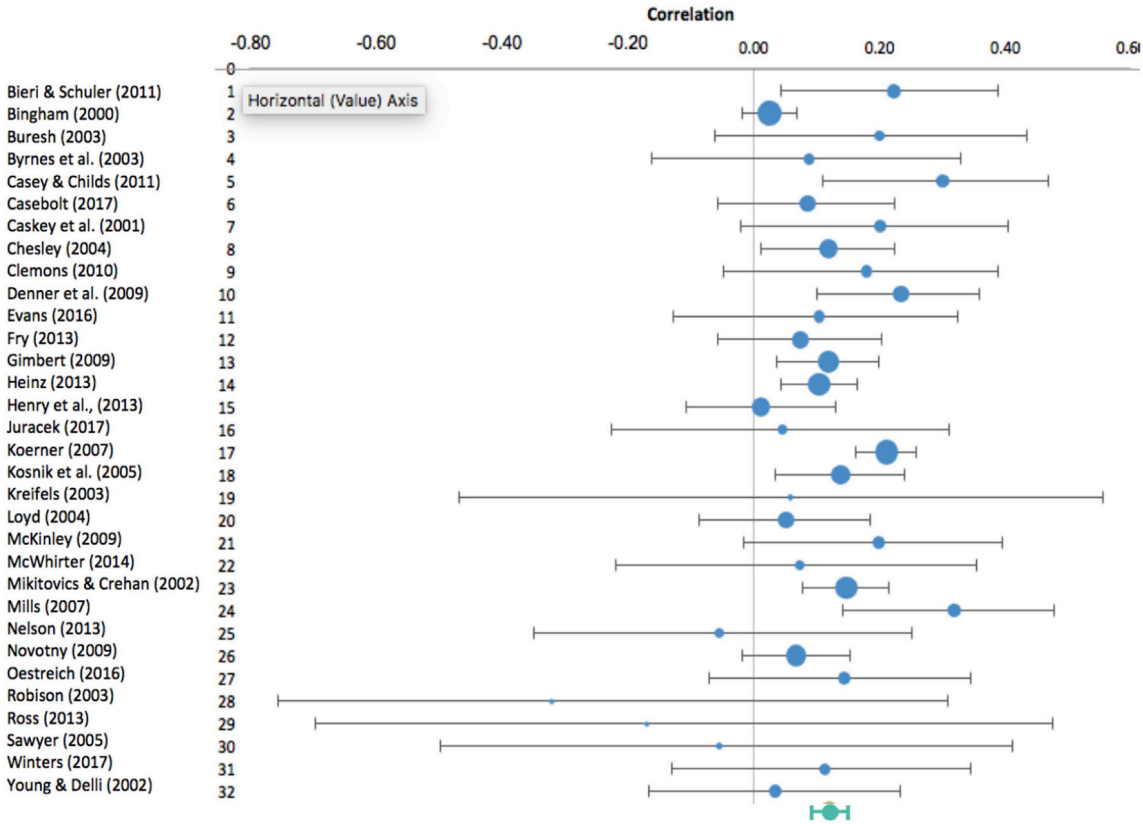

Fig. 4. Synthesis forest plot for all included studies. Bracketed lines with blue circles indicate 95% confidence interval for each study and the bracketed line with a green circle indicate the overall 95% confidence interval and the overall effect size. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Sub-Group Analysis: Number of Effect Sizes, Mean r, 95% Confidence Intervals, and Heterogeneity Indicators.

Variable	*k*	*r*	95% CI (LL, UL)	*Q*	p_Q	I^2 (%)
Overall teacher effectiveness	32	.12	(.10, .15)	55.45	.00	44.09
Moderator analysis						
Selection method type	37	.11	(.08,.14)	68.33	.00	47.31
Academic	15	.13	(.07, .19)	35.61	.00	60.68
Non-academic	22	.10	(.07, .13)	24.32	.28	13.65
Teacher effectiveness measure	37	.12	(.08, .16)	74.71	.00	51.81
Performance evaluation	20	.13	(.09, .17)	21.47	.31	11.52
Student Academic Achievement	12	.08	(-.01, .17)	41.51	.00	73.50
Teacher Academic Achievement	3	.13	(.06, .19)	0.96	.62	.00
Absenteeism	2	.09	(-1.00, 1.00)	9.65	.00	89.64
Educational level	35	.12	(.11, .13)	130.79	.00	74.00
Elementary	10	.12	(-.03, .26)	85.35	.00	89.46
Secondary	5	.07	(-.14, .29)	16.63	.00	75.94
Mixed	20	.12	(.10, .15)	10.91	.93	0.00
Selection type	32	.12	(.10, .15)	55.86	.00	44.50
ITE	9	.14	(.08, .20)	12.09	.15	33.85
Employment	23	.11	(.07, .15)	42.73	.01	48.52
Output type	32	.12	(.10, .15)	55.86	.00	44.50
Journal	11	.13	(.09, .18)	50.14	.22	23.44
Dissertation	21	.11	(.06, .15)	42.15	.00	52.55

Note. *k* = number of effect sizes, 95% CI (LL, UL) = 95% Confidence Interval (Lower Limit, Upper Limit), *Q* = *Q*-Statistic, p_Q = *p*-value for the *Q*-statistic.

confidence interval ranging from 0.10 to 0.15. A synthesis of individual studies with its weighted point estimates is displayed in Fig. 4.

We used two measures of heterogeneity: *Q-statistic* (or Cochrane's Q) with its *p*-value, and I^2. The *Q*-statistic was significant, $Q(31) = 55.45$, $p < .001$, indicating that there are significantly different effect sizes across studies. Furthermore, I^2 was 44.09% indicating that there was 'low-to-moderate' variability across the studies (using the descriptor provided by Higgins, Thompson, Deeks, & Altman, 2003), hence indicating the suitability of moderator analyses (Borenstein, Hedges, Higgins, & Rothstein, 2009). Heterogeneity statistics are reported in Table 2.

Three indicators were used to assess publication bias in the estimates using a fixed estimates model. First, the relatively symmetrical spread of the estimates in the funnel plot (Fig. 5) indicates that there may not be a publication bias. Second, the significant Egger's regression test (Egger, Davey, Schneider, & Minder, 1997) with an intercept estimate of −0.10 (95% CI: −0.96, 0.77) is another indicator that there may not be a publication bias. Finally, Rosenthal's (1979) fail-safe *N* tests indicate that a large number of studies (i.e., 963 studies) with non-significant results would be needed to make the association non-significant, which again indicates that there may not be a publication bias.

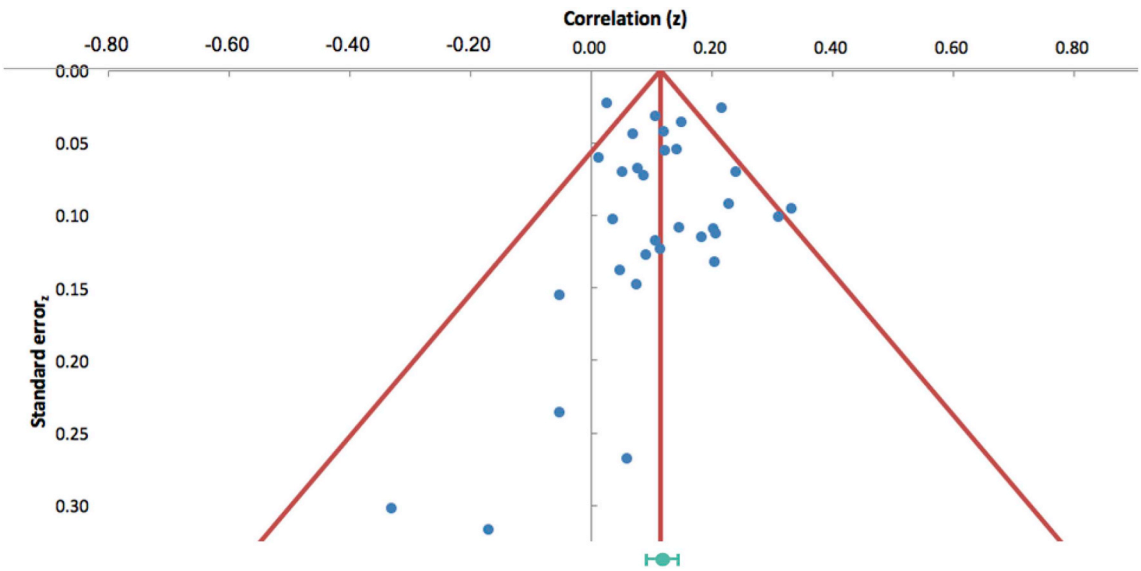

Fig. 5. Funnel plot with estimates from the included studies with standard error in the y axis and effect size on the x axis.

3.2.1. Variation in effect sizes

As can be seen in the synthesis forest plot (Fig. 4), 28 studies showed positive effect sizes and 4 showed negative effect sizes. Ten studies reported significant findings, all positive. The range of effect sizes and statistical significance is likely due to multiple factors: restricted variance in the independent and dependent variables, uncertain reliability and validity of the measures used, and low sample sizes. The ten studies that showed significantly positive effect sizes comprised studies with academic predictors ($n = 3$), non-academic predictors ($n = 3$), and combined academic and non-academic predictors ($n = 4$). In contrast to the publication source reported for the 32 studies in the overall meta-analysis, where 66% of the studies were dissertations, the majority (70%) of the studies with significant positive effect sizes were journal publications.

Four studies had negative effect sizes. Robison (2003) examined the differences in student academic growth for teachers who, on the Venture for Excellence interview (VFE), scored high (14 or above) and low (12 or below). However, the distribution and the range of scores within the high and low scoring groups was not provided. Given the small sample size (14 participants), the distribution of the VFE scores may be small within each group. Furthermore, the mean differences in the student academic growth scores were small (e.g., 0.46 and 2.99), with large standard errors (1.68 and 1.06). Ross (2013) examined the correlation between interview scores and student academic achievement. The range of dispositional interview scores was small (12–22) as were the student academic achievement scores (e.g., 380 to 450). With the small sample size of 7–15 participants, the uneven spread of the scores may have skewed the results. Given the wide range of studies reported in the meta-analysis, especially in terms of sample size, evidence of validity and reliability, and scoring range, one must be cautious when interpreting individual studies within the meta-analysis.

3.3. Moderator analyses

Moderator analyses were conducted to further examine the relation between teacher selection measures and teacher effectiveness measures. To start, we tested for publication bias by testing any differences in effect size by publication category (i.e., journal articles or dissertations). The four main analyses examined: (a) type of predictor (academic or non-academic), (b) type of teacher effectiveness measure, (c) teaching level (elementary or secondary), and (d) selection type (for employment or for entry into ITE program).

To compute the overall effect size for each moderator analysis, individual Pearson *r*s were transformed into Fisher's *z* scores, which were then averaged within each study by the type of moderator analysis category. Fisher's *z* scores were averaged because including multiple effect sizes from the same study would violate the assumptions of independence (Lipsey & Wilson, 2001). For example, in a moderator analysis separated by type of predictor, a study containing three effect sizes for academic predictors and two effect sizes for non-academic, the predictors were averaged such that there was only one effect size for the academic predictor type and one correlation for non-academic predictor. The average Fisher's *z* scores were then transformed back to Pearson *r*s. These transformed Pearson *r*s were then used to calculate the final effect sizes for the meta-analysis.[2] Table 2 contains the type of moderator analysis conducted with the corresponding number of effect sizes (k), the actual effect sizes, and the 95% lower and upper confidence intervals.

3.3.1. Moderation by publication category ($k = 32$)

The associations between the predictors and outcomes were significant when the results were published in journals ($r = 0.13$, $df = 10$, $p < .05$), and in dissertations ($r = 0.11$, $df = 20$, $p < .05$), and there were no significant differences in the strength of the associations, Q_b (1) = 0.64, p = ns.

3.3.2. Moderation by type of predictor ($k = 37$)

The associations with the outcomes were significant for both academic ($r = 0.13$, $df = 14$, $p < .05$) and non-academic predictors ($r = 0.10$, $df = 21$, $p < .05$), and the two associations were significantly different from each other, Q_b (1) = 8.40, $p < .05$.

3.3.3. Moderation by type of teacher effectiveness measure ($k = 37$)

The predictors were significantly associated with performance evaluations ($r = 0.13$, $df = 19$, $p < .05$) and practicum ratings, ($r = 0.13$, $df = 2$, $p < .05$), but not with student academic performance ($r = 0.08$, $df = 11$, $p = ns$), or absenteeism ($r = 0.09$, $p = ns$). The four categories did not differ significantly from each other, Q_b (3) = 1.11, $p > .05$.

3.3.4. Moderation by selection purpose ($k = 32$)

The associations between the predictors and outcomes were significant both when used for the purposes of selection for employment ($r = 0.11$, $df = 22$, $p < .05$), and for ITE admissions ($r = 0.14$, $df = 8$, $p < .05$), with no significant difference in the strength of the associations, Q_b (1) = 1.03, p = ns.

3.3.5. Moderation by educational level ($k = 35$)

The associations between the predictors and outcomes were only significant when at mixed primary and secondary levels ($r = 0.12$, $df = 19$, $p < .05$), and not when they were examined at each educational level. Again, there were significant differences in

[2] An alternative method to not violate the assumption of independence would be to use robust variance estimation in the calculation (Sidik & Jonkman, 2006).

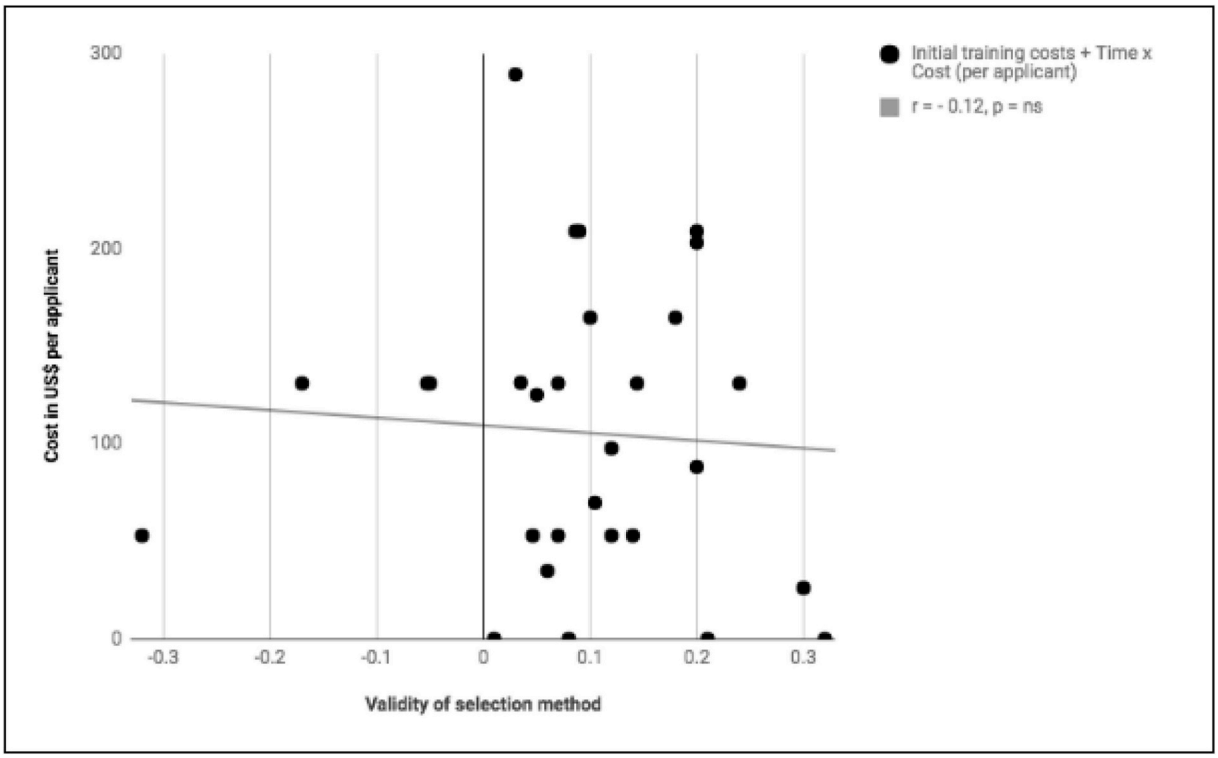

Fig. 6. Correlation between validity and cost of selection method.

the associations between the categories, Q_b (2) = 17.91, $p < .05$.

3.4. Cost of selection methods

The mean cost of selection methods for the receiving institution or school district was US $104.32 per candidate, with a low of $0 in four studies where existing data such as secondary school or university grades were used for selection (Heinz, 2013; Henry et al., 2013), or where required assessments were used with costs borne by the candidates (e.g., Denner, Norman, & Lin, 2009). The highest costs (e.g., $289.61, $209.13) were calculated for assessment centers requiring multiple assessors (e.g., Bieri & Schuler, 2011) and selection methods that required initial training costs (e.g., $2500 for the TPI in Bingham, 2000) in addition to time required for interviews. Cost estimates should be interpreted cautiously: the annual costs for selection methods involving initial training costs would decrease over time, and the number of personnel used for selection likely varies across sites. Fig. 6 shows the bivariate relations between the validity and cost of selection methods for the 32 studies in the review. The overall correlation between predictive validity and cost of implementation of selection method was negative but non-significant, ($r = -0.12$, p = ns).

3.5. Fairness and bias

Some selection methods have been shown to be prone to inter-group differences, and may inadvertently favor some groups (race, gender, SES) over others (Shultz & Zedeck, 2012). We searched the 32 articles for any discussion of inter-group differences or potential bias in results. Fry (2013) discussed the higher Teacher Insight scores of male math teachers; Loyd (2004) noted no significant differences for race in scores on the Urban Teacher Perceiver Interview; McKinley (2009) noted gender differences on scores on the Ventures for Excellence screening tool; and Mikitovics and Crehan (2002) noted that the “supply of minority teacher candidates appears to be negatively affected by the use of the Pre-Professional Skills Test as a screening device” (p. 221). No other studies addressed fairness and potential biases of selection methods.

4. Discussion

The purpose of this article was to review research on teacher and prospective teacher selection methods, to evaluate the predictive validity of these methods, and to consider a model for teacher selection that reflects research and theoretical work conducted in and outside of education. Our findings confirm that selection methods assessing academic and non-academic constructs are statistically associated with teacher effectiveness measures, although the effect sizes were small. The effect size for non-academic predictors ($r = 0.10$) was significantly smaller than for academic predictors ($r = 0.14$), reflecting the particular challenge of reliably and validly

evaluating non-academic attributes in high-stakes settings (e.g., Sackett et al., 2017). Methods for selection into employment and ITE both significantly predicted subsequent teacher effectiveness, but again, the effect sizes were small. There was little indication of publication bias, with significant selection method/effectiveness associations in journal articles and in unpublished dissertations, and with publication bias indices (e.g., Egger's regression test) suggesting against publication bias.

The relative weakness of predictive validity in comparison with other disciplines raises questions about methods used for selection in education. The findings showed that the best (single) estimate of the relation between teacher selection methods and teacher effectiveness is $r = 0.12$ (considered a *small* effect size according to Hattie, 2009), with both academic and non-academic predictors showing significant but small effect sizes. Of the 32 studies, 28 studies showed a positive relation between selection method and teacher effectiveness, but only 10 studies had overall effect sizes that were significantly different from 0.

The predictive validities of selection methods in this review are generally lower than those found in research in other fields. The discrepancy in predictive validity between education and other fields may be due to the fact that teaching is uniquely complex compared to other professions, and thus more difficult to predict, or it may be that education has benefited less from the extensive research and development that has characterized other fields (e.g., Ryan & Ployhart, 2014; Schmidt & Hunter, 1998). Teaching is undoubtedly multi-faceted and complex, but Harris and Rutledge (2010) point out that numerous professions offer similar levels of complexity.

In addition, we found a weaker relation between predictors and teaching outcomes than did Metzger & Wu in their 2008 review of the TPI, where they reported a mean correlation of $r = 0.28$ between the TPI and teacher effectiveness. There are several explanations for this discrepancy. First, Metzger and Wu's studies consisted largely of dissertations and test publisher reports, with only 2 of 24 studies coming from peer-reviewed journals. We chose not to include non-peer reviewed data from test publishers, but our comparison of data sources showed a significant (but modest) difference in effect size favoring dissertations over journal articles. Second, in half of the studies in the Metzger and Wu article, the TPI was administered after hiring; that is, it was not used for high-stakes selection. The correlation between the TPI administered in a low-stakes research setting ($r = 0.35$) was significantly higher than when used for actual selection ($r = 0.21$). Third, Metzger and Wu studies were pre-2000 and ours were post-2000, and it could be argued that teaching is becoming increasingly complex, and increasingly difficult to capture as an outcome.

Another factor to consider when examining the correlation between teacher selection and teacher effectiveness is the quality (reliability and validity) of the measures of teacher effectiveness. In our meta-analysis, most of the outcome measures consisted of teaching observation ratings completed by a principal or supervisor using a wide range of scoring protocols, but with little evidence supporting their reliability and validity. Our results showed a significant relationship between selection methods and performance evaluations, but not for student academic performance or for student teachers' practicum ratings. The need for developing improved measures of teacher effectiveness has been the focus of recent funding initiatives, but the challenges remain considerable (e.g., van der Lans et al., 2018). Finding and adopting valid methods of measuring teacher effectiveness presents a real challenge for practitioners (e.g., Patrick & Mantzicopoulos, 2016; Strong, Gargani, & Hacifazlioglu, 2011). Recent efforts to improve teacher evaluation systems (e.g., Gill et al., 2016) have led to more robust measures such as the Classroom Assessment Scoring System (CLASS; Hamre et al., 2013) and the International Comparative Analysis of Learning and Teaching (ICALT; van der Lans et al., 2018), but the studies included in this review did not use these more recent, evidence-supported observation methods. Using less reliable and valid outcome measures weakens the correlation between predictor and criterion.

A key question when assessing predictive validity is the practical value of the effect sizes found in the studies: do current selection methods make a difference in educational outcomes? Ferguson (2009) lays out a lower limit of $r = 0.20$ for practical significance in social science research, a standard echoed by Hedges and Hedberg (2007) vis-à-vis educational research when objective measures of outcomes are used. However, teacher selection methods are often applied to large numbers of applicants, and Coe (2002) points out that small effect sizes can make a big difference to educational outcomes if the effect is applied to a large sample.

4.1. Selecting teachers and prospective teachers

For selection into training or employment, selection methods are most likely to be useful when the number of predictions is large and when effects are additive; for example, small but significant effect sizes for academic and non-academic selection methods provide added value—incremental validity—over and above the effects shown by the use of only one type of predictor. In medicine, for example, tests of non-academic attributes explain a significant amount of explained variance over and above academic tests (e.g., Patterson et al., 2016). Although there is not much predictive gain for any particular selection method for a specific selection decision, small gains in predictive accuracy lead to important improvements at the organizational level (Morgesen et al., 2007). The evidence from the current review suggests that academic and non-academic selection methods both show potential for contributing to selection accuracy for employment and ITE programs, but that further attention to research and development is needed, especially to develop reliable and valid non-academic selection methods.

4.1.1. Selection into employment and ITE programs

The methods for selection into employment (r = 0.11) and into ITE programs (r = 0.14) were significantly related with teacher effectiveness outcomes. Many of the attributes relevant for selecting teachers for employment are qualitatively similar to the attributes relevant for selection into ITE programs, because key personal characteristics are relatively stable over time [Steyer et al., 1999]), notwithstanding the importance of developmental interactions with contextual factors such as learning opportunities and teaching experience as laid out in Kunter's COACTIVE model. Lievens and Sackett (2012) finding that the attributes measured at admission into professional training are predictive of professional competence up to nine years later supports the notion that personal

characteristics measured at selection into training interact with learning opportunities to predict effectiveness not just during initial training but long into professional practice.

4.1.2. Cost of selection methods

We did not find a positive relation between the cost of selection methods (M = $104.32) and their predictive validity; in fact, the correlation trended negative ($r = -0.12$, p = ns). Selection methods can be designed with great care, but the ratio of effectiveness, benefit, or utility to cost needs to be considered, especially in large scale selection systems. In some cases, personnel costs can be very high: in one study we reviewed, but that did not meet the inclusion criteria for the meta-analysis (Denner et al., 2001), a comprehensive selection method for ITE was implemented involving a biographical essay, a statement of informed beliefs, a written profile of a K-12 student, and a case study analysis, all of which were scored by a team of faculty members. Then, a personal interview was conducted by "a collaboration of arts and sciences faculty, education faculty, undergraduate students, and professionals from the field" (p. 165). Further selection indicators including instructor checklists and statement on candidate suitability were collected. The entire process resulted in 96% (181/188) of candidates proceeding to interview and subsequently being offered admission to the program. Our intention in spotlighting this example is not to expose poor practice—there may be good reasons to collect these data, and de-selecting 7 candidates from 188 may be worth the cost—but we suggest that evaluating the cost-benefit ratio of selection methods is a worthwhile practice for any ITE program or hiring school district.

Although it is relatively straightforward to calculate the cost of using a selection method, or to link cost with predictive validity, it is more difficult to estimate the 'return on selection,' or the value of using a particular approach to selection. One framework to understand the use of decision-making in personnel selection is utility analysis, reflecting a family of theories and methods which are designed to describe and predict the usefulness of personnel decisions, such as selection (Sturman, 2012). Although placing an economic value on teacher effects is controversial (e.g., Paufler & Amrein-Beardsley, 2014), the approach may be useful in future research to calculate the relative costs and benefits of teacher selection methods.

4.1.3. Fairness and bias of selection methods

Few of the reviewed studies attended to issues of fairness and bias in selection methods, with scant mention of inter-group differences or potential bias. This is an important omission: in the United States, the proportion of non-White students far outweighs the number of non-White teachers (White, DeAngelis, & Lichtenberger, 2013), and researchers have noted the positive impact of racial/ethnic minority teachers on minority student academic and social outcomes (e.g., Villegas & Irvine, 2010). Adverse impact can result from systematic bias in personnel selection methods, whereby members of a sub-group are selected more or less often than a member of another sub-group. Concerns about systematic bias in selection methods that focus on academic attributes have undermined the attraction of some traditional selection measures in other fields (e.g., Kidder, 2001; Lievens, Patterson, Corstjens, Martin, & Nicholson, 2016). Our review suggests that few researchers are examining potential bias in the application of teacher selection methods, allowing the potential for adverse impacts (e.g., difference in selection rates between different subgroups).

4.2. Learning from selection research in other disciplines

We suggest that important lessons can be learned from selection research in other disciplines. The 'more effective' methods to assess non-academic attributes listed in Patterson et al.'s, (2016) review of selection methods for medical training (i.e., SJTs and MMIs) were not found in the studies we reviewed. Assessment of academic attributes through evaluation of academic records (deemed an effective practice by Patterson et al.) was included in the selection methods in most studies we reviewed, but explicit evaluation of reasoning ability was not, in spite of findings from organizational psychology that reasoning ability is one of the best predictors of training and workplace outcomes (Bowles et al., 2014; Schmidt & Hunter, 1998).

Organizational psychologists design selection systems to answer questions about predictor variables, such as *What should be assessed?* and *How should we assess it?* but attention is also paid to assessing a broad range of outcome measures over longer periods of time (Beier & Ackerman, 2012). The relation between predictor and criterion can be non-linear, with key predictor variables waxing and waning in predictive power over time (e.g., Lievens & Patterson, 2011; Marks & Moss, 2016). In our review, teacher effectiveness was primarily defined as observed teaching performance and student achievement gains measured at one time point. In organizational psychology, productivity and effectiveness change as workers sharpen skills, move into leadership positions, and reorganize their goals and redistribute their energy in the workplace. We know very little about how academic and non-academic variables assessed at the point of selection predict teacher effectiveness at different career stages. Research designs that use hierarchical methods can be used to simultaneously examine within- and between-person change over time, thus allowing the testing of models that link attributes measured at selection with various facets of teacher effectiveness over time. For example, tracing the development of the predictive validity of non-academic attributes of personality and motivation over time would allow for new insight into selection methods, but would also increase our understanding about how important teaching-related attributes change over the course of a career.

4.3. Considerations for practitioners

For practitioners, the results from our study suggest that current selection methods provide modest predictive utility for selecting prospective teachers, and that the costs of selection practices are not associated with increases in predictive validity. The selection methods reviewed in the study did not reflect advances in other professional disciplines, and the outcome measures of teacher

effectiveness did not tend to reflect recent advances in educational research. We propose four ways that education practitioners can improve their selection methods: (a) design and test selection methods that include academic and non-academic measures that build on relevant theories from education and psychology, (b) consider adopting selection methods from other professional disciplines that have shown evidence of reliability and validity (e.g., SJTs and MMIs), (c) adopt state-of-the-art methods to assess teacher effectiveness, with evidence of reliability and validity (e.g., Hamre et al., 2013), and (d) routinely evaluate selection methods by considering the cost-effectiveness of current measures, and by collecting data linking selection methods and teaching outcomes (teaching behaviors, retention rates) gathered after selection.

4.4. Research agenda

Designing effective and efficient selection methods for teaching jobs and entrance into ITE programs presents many challenges, but these challenges have been successfully addressed in other professional fields. Our results support government calls (e.g., U.K. House of Commons, 2012) to improve teacher selection methods, and especially the methods used to assess non-academic attributes, which we found to be significantly weaker than tests of academic attributes. We propose that researchers develop selection methods that are grounded in educational theory and research, and that reflect recent advances in organizational psychology and medical education. Two recent programs of research, in Australia (e.g., Bowles et al., 2014) and in the U. K. (Klassen et al., 2017; Klassen & Kim, 2018), are focused on building theory- and evidence-based teacher selection methods; however, more work is needed in order to establish a robust evidence base. In parallel with the aforementioned considerations for practitioners, we offer five recommendations for researchers: (a) develop and test selection methods in line with educational and psychological theory and building on research from other professional fields, (b) develop and test better measures of non-academic attributes at the point of selection, (c) investigate the longitudinal relationship between teacher characteristics—measured at the point of selection—and teacher effectiveness using recent validated measures (Gill et al., 2016), (d) investigate the fairness of current and future selection methods vis-à-vis under-represented ethnic and other groups not well represented in the teaching workforce, (e) increase inter-disciplinary cross-talk between education and other fields (especially organizational psychology and medical education) where selection research has a longer history and a more robust evidence base.

4.5. Limitations

The generalizability of the findings from this review is limited by the paucity of research that examines teacher selection tools administered at the point of selection. Research on constructs associated with teacher effectiveness is common (e.g., research on teacher motivation), but comparatively few studies have assessed these constructs in high-stakes selection situations. The results from this study are dependent on the characteristics of the applicants who self-selected for teacher training and employment, and may not represent the population of practicing teachers. In addition, the results are restricted by the limitations of the search procedures that included English-language only studies, and by relevant articles that may have been missed in the search procedures. Although most of the data coding decisions were unanimous, or near-unanimous, the inclusion of an inter-rater reliability statistic (e.g., Cohen's kappa) would have provided additional support for the coding decisions. Many of the studies used teacher effectiveness outcome variables that lacked evidence of their validity (see Gill et al., 2016, and Strong et al., 2011 for discussion of teacher observation methods). The estimated costs for selection methods should be interpreted cautiously, since the estimates are based on a number of assumptions about rates of pay and number of personnel.

The studies included in the meta-analysis used a variety of outcome variables, ranging from state-mandated measures (e.g., Northwest Evaluation Association reading growth score) to researcher-created scales (e.g., Teacher performance rating scale). Given the range, it was difficult to standardize these measures; furthermore, very little evidence regarding reliability and validity was provided for these outcome measures. As more research is conducted to build evidence on teacher effectiveness measures, we expect that more meaningful effect size calculations can be made in the future.

A limited dataset prevented examination of moderators such as gender, ethnicity, and socio-economic status that were of interest for the purpose of the review. Unfortunately, few studies included sufficient information necessary to calculate these effects. Calculation of effect sizes were limited by the use of study-level instead of outcome-level analytic procedures. Although these procedures met the independence assumption required in most meta-analyses, some data were lost within the study due to aggregating multiple outcomes. Finally, our decision to exclude reports from commercial test providers means that our findings do not cover the whole spectrum of research produced on selection methods.

5. Conclusion

Our review of the teacher selection literature reveals some support for the use of academic and non-academic predictors of teacher success, both for employment and entry into ITE, but the evidence base is not very robust, and there is an over-reliance on cross-sectional studies. We found no longitudinal research exploring the relation between selection methods and how teacher effectiveness develops over time. In addition, there is little consideration of how important theoretical foundations can underpin the development of selection methods. We expect that with a continued research and policy focus on ways to improve teacher effectiveness, increased attention will be paid to the ways in which we select prospective teachers.

Acknowledgments

This work was supported by the European Research Council (Consolidator Grant no. 647234; project acronym SELECTION) awarded to the first author.

References

Abdelfattah, F. (2010). The relationship between motivation and achievement in low-stakes examinations. *Social Behavior and Personality: International Journal, 38*, 159–167.
Antoniou, P., Kyriakides, L., & Creemers, B. P. M. (2015). The dynamic integrated approach to teacher professional development: Rationale and main characteristics. *Teacher Development, 19*, 535–552.
Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Predicting early career teacher effectiveness. *AERA Open, 1*(4), 1–23.
Bahr, N., & Mellor, S. (2016). *Building quality in teaching and teacher education.* Camberwell, Australia: ACER Press.
Bangerter, A., Konig, C. J., Blatti, S., & Salvisberg, A. (2009). How widespread is graphology in personnel selection practice? A case study of a job market myth. *International Journal of Selection and Assessment, 17*, 219–230.
Barber, M., & Mourshed, M. (2007). *How the world's best performing school systems come out on top.* London: McKinsey & Company.
Beier, M. E., & Ackerman, P. L. (2012). Time in personnel selection. In N. Schmitt (Ed.). *The Oxford handbook of personnel assessment and selection* (pp. 721–739). Oxford: Oxford University Press.
*Bieri, C., & Schuler, P. (2011). Cross-curricular competencies of student teachers: A selection model based on assessment centre admission tests and study success after the first year of teacher training. *Assessment & Evaluation in Higher Education, 36*, 399–415.
*Bingham, P. J. (2000). *The concurrent and predictive validity of elementary school teacher pre-employment success indicators (Unpublished doctoral dissertation)*Austin, Texas: The University of Texas.
Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic approaches to a successful literature review* (2nd ed.). London: Sage.
Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Chichester: John Wiley & Sons, Ltd.
Bowles, T., Hattie, J., Dinham, S., Scull, J., & Clinton, J. (2014). Proposing a comprehensive model for identifying teaching candidates. *Australian Educational Researcher, 41*, 365–380.
*Buresh, R. J. (2003). *The predictive validity of the Teacher Perceiver Interview in selecting effective elementary teachers in a mid-sized midwestern school district (Unpublished doctoral dissertation)*Grand Forks, North Dakota: University of North Dakota.
Byrnes, D. A., Kiger, G., & Shechtman, Z. (2003). Evaluating the use of group interviews to select students into teacher-education programs. *Journal of Teacher Education, 54*, 163–172.
*Casebolt, P. C. (2016). *Teacher hiring explored: Examining the effectiveness of educator selection processes*(Unpublished doctoral dissertation), 2016Nashville, TN: Trevecca Nazarene University.
*Casey, C., & Childs, R. (2011). Teacher education admission criteria as measure of preparedness for teaching. *Canadian Journal of Education, 34*, 3–20.
Casey, C., & Childs, R. A. (2017). Teacher education program admission criteria and what beginning teachers need to know to be successful teachers. *Canadian Journal of Educational Administration and Policy, 67*, 1–24.
*Caskey, M. M., Peterson, K. D., & Temple, J. B. (2001). Complex admission selection procedures for a graduate preservice teacher education program. *Teacher Education Quarterly, 28*(4), 7–21.
Chamorro-Premuzic, T., & Furnham, A. (2010). *The psychology of personnel selection.* New York: Cambridge University Press.
*Chesley, D. (2004). *The validity of teacher employment success predictors (Unpublished doctoral dissertation)*Washington, D. C: George Washington University.
Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *The American Economic Review, 104*, 2593–2632.
Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83–117.
*Clemons, B. L. (2010). *A correlational study of the Ventures for Excellence Interview-rating system and first year teacher evaluations (Unpublished doctoral dissertation)*St. Charles, MO: Lindenwood University.
Coe, R. (2002). September). It's the effect size, stupid: What effect size is and why it is important. *Paper presented at the annual conference of the British educational research association* (Exeter, England).
Cook, M. (2009). *Personnel selection: Adding value through people* (5th ed.). Chichester, UK: Wiley-Blackwell.
Dana, J., Dawes, R., & Peterson, N. (2013). Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making, 8*, 512–520.
Darling-Hammond, L. (2010). Teacher education and the American future. *Journal of Teacher Education, 61*, 35–47.
Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teachers:" what does "scientifically-based research" actually tell us? *Educational Researcher, 31*, 13–25.
Davies, P., Connolly, M., Nelson, J., Hulme, M., Kirkman, J., & Greenway, C. (2016). 'Letting the right one in': Provider contexts for recruitment to initial teacher education in the United Kingdom. *Teaching and Teacher Education, 60*, 291–302.
*Denner, P., Norman, A., & Lin, S. Y. (2009). Fairness and consequential validity of teacher work samples. *Educational Assessment, Evaluation and Accountability, 21*, 235–254.
Denner, P. R., Salzman, S. A., & Newsome, J. D. (2001). Selecting the qualified: A standards-based teacher education admission process. *Journal of Personnel Evaluation in Education, 15*, 165–180.
Ebmeier, H., Dillon, A., & Ng, J. (2007, October). Employment selection instruments—what we have learned from ten years of research. *A paper presented at the annual conference of the American Association of School Personnel Administrators* (Kansas City, KS).
Ebmeier, H., & Ng, J. (2005). Development and field test of an employment selection instrument for teachers in urban school districts. *Journal of Personal Evaluation in Education, 18*, 201–218.
Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629–634.
*Evans, K. E. (2016). *The role of teacher TeacherInsight scores and teacher demographic characteristics in the identification of effective teachers: Using student performance as a validation tool.*
Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*, 532–538.
*Fry, T., Jr. (2013). *Exploring the relationship between the TeacherInsight score and the Teacher Growth Index (Unpublished doctoral dissertation)*Ashland, Ohio: Ashland University.
Gallup Organization (2001). *Teacher Perceiver predictive validity study.* Lincoln, NE: Author.
Gayton, S. D., & Kehoe, J. E. (2015). A prospective study of character strengths as predictors of selection into the Australian Army Special Force. *Military Medicine, 180*, 151–157.
Gill, B., Shoji, M., Coen, T., & Place, K. (2016). *The content, predictive power, and potential bias in widely used teacher observation instruments (REL 2017–191).* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Regional Educational Laboratory Mid- Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs.
*Gimbert, B. G., & Chesley, D. (2009). Predicting teacher success using teacher selection practices and classroom performance assessment. *Journal of School Leadership, 19*, 49–80.
Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher, 44*, 87–95.
Goldhaber, D., Grout, C., & Huntington-Klein, N. (2014). *It's selective but is it effective? Exploring the predictive validity of teacher selection tools. CEDR policy brief 2014-9.* Seattle, WA: University of Washington.
Gottheil, E., & Michael, C. M. (1957). Predictor variables employed in research on the selection of medical students. *Academic Medicine, 32*, 131–147.
Greenberg, J., McKee, A., & Walsh, K. (2015). *Teacher prep review: A review of the nation's teacher preparation programs.* National Council on Teacher Quality. Retrieved

from http://www.nctq.org/dmsView/Teacher_Prep_Review_2014_Report.
Grissom, J. A., & Loeb, S. (2017). Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low- and high-stakes environments. *Education Finance and Policy, 12*, 369–395.
Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., ... Hamagami, A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal, 113*, 461–487.
Hanushek, E. A. (2014). Boosting teacher effectiveness. In C. E. J. Finn, & R. Sousa (Eds.). *What lies ahead for America's children and their schools* (pp. 23–35). Stanford, CA: Hoover Institution Press.
Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics, 4*, 131–157.
Harris, D. N., & Rutledge, S. A. (2010). Models and predictors of teacher effectiveness: A comparison of research about teaching and other occupations. *Teachers College Record, 112*, 914–960.
Hattie, J. A. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Oxford: Routledge.
Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60–87.
*Heinz, M. (2013). Tomorrow's teachers—selecting the best: An exploration of the quality rationale behind academic and experiential selection criteria for initial teacher education programmes. *Educational Assessment, Evaluation and Accountability, 25*, 93–114.
*Henry, G. T., Campbell, S. L., Thompson, C. L., Patriarca, L. A., Luterbach, K. J., & Lys, D. B. (2013). The predictive validity of measures of teacher candidate programs and performance: Toward an evidence-based approach to teacher preparation. *Journal of Teacher Education, 64*, 439–453.
Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557–560.
Hindman, J., & Stronge, J. (2009). The $2 million decision: Teacher selection and principals' interviewing practice. *Educational Research Service Spectrum, 27*, 1–10. Retrieved from http://humanresourcesportfolio.wikispaces.com/.
Hobson, A. J., Ashby, P., McIntyre, J., & Malderez, A. (2010). *International approaches to teacher selection and recruitment. OECD education working papers.* Paris: OECD Publishing No. 47.
Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics, 6*, 801–825.
*Juracek, T. (2016). *Predictability of hiring mechanisms for elementary teachers*Unpublished doctoral dissertationOmaha, NE: University of Nebraska.
Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Seattle, WA: Bill and Melinda Gates Foundation.
Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation.
Kausel, E. E., Culbertson, S. S., & Madrid, H. P. (2016). Overconfidence in personnel selection: When and why unstructured interview information can hurt hiring decisions. *Organizational Behavior and Human Decision Processes, 137*, 27–44.
Kidder, W. C. (2001). Does the LSAT mirror or magnify racial and ethnic differences in educational attainment? A study of equally achieving" elite" college students. *California Law Review, 1055*, 1074–1124.
Klassen, R. M., & Dolan, R. (2015, September). Selection for teacher education in the UK and the Republic of Ireland: A proposal for innovation. *Presented at the meeting of the European Conference on Educational Research, Budapest, Hungary*.
Klassen, R. M., & Durksen, T. L. (2014). Weekly self-efficacy and work stress during the final teaching practicum: A mixed methods study. *Learning and Instruction, 33*, 158–169.
Klassen, R. M., Durksen, T. L., Kim, L. E., Patterson, F., Rowett, E., & Warwick, J. (2017). Developing a proof-of-concept selection test for entry into primary teacher education programs. *International Journal of Assessment Tools in Education, 4*, 96–114.
Klassen, R. M., & Kim, L. E. (2018). Development of an online construct-informed situational judgment test for screening applicants for initial teacher education. TSP Working Paper 2018-2 https://doi.org/10.31234/osf.io/64hry.
Klassen, R. M., & Tze, V. M. C. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review, 12*, 59–76.
*Koerner, R. J. (2007). *The relationship between the TeacherInsight score and student performance as measured by the Texas Growth Index (Unpublished doctoral dissertation)* Denton, Texas: University of North Texas.
*Kosnik, C., Brown, R., & Beck, C. (2005). The preservice admissions process: What qualities do future teachers need and how can they be identified in applicants? *The New Educator, 1*, 101–123.
*Kreifels, R. G. (2003). *A study of selected teacher interview characteristics and student achievement (Unpublished doctoral dissertation)*St. Louis, MO: Saint Louis University.
Kunter, M., Kleickmann, T., Klusmann, U., & Richter, D. (2013). The development of teachers' professional competence. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.). *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 63–77). New York: Springer U.S.
Kuykendall, L., Tay, L., & Ng, V. (2015). Leisure engagement and subjective well-being: A meta-analysis. *Psychological Bulletin, 141*, 364–403.
van der Lans, R. M., van de Grift, W. J., & van Veen, K. (2018). Developing an instrument for teacher feedback: Using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. *The Journal of Experimental Education, 86*, 247–264.
Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology, 96*, 927–940.
Lievens, F., Patterson, F., Corstjens, J., Martin, S., & Nicholson, S. (2016). Widening access in selection using situational judgment tests: Evidence from the UKCAT. *Medical Education, 50*, 624–636.
Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology, 97*, 460–468.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.
Liu, E., & Johnson, S. M. (2006). New teachers' experiences of hiring: Late, rushed, and information poor. *Educational Administration Quarterly, 42*, 324–360.
*Loyd, D. E. (2004). *The effect of a screening instrument to determine teacher effectiveness as measured by academic gain scores (Unpublished doctoral dissertation)*Knoxville, TN: The University of Tennessee.
Marks, A. B., & Moss, S. A. (2016). What predicts law student success? A longitudinal study correlating law student applicant data and law school outcomes. *Journal of Empirical Legal Studies, 13*, 205–265.
McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740.
*McKinley, K. J. (2009). *Exploring the efficiency and effectiveness of teacher selection tools: The effects on the total group with a focus on the experience sub-groups (Unpublished doctoral dissertation)*St. Charles, MO: Lindenwood University.
*McWhirter, A. B. (2014). *The correlation between a teacher selection instrument and teacher quality in Tennessee (Unpublished doctoral dissertation)*Phoenix, AZ: Grand Canyon University.
Mertz, N. T. (2010). Teacher selection and school leader effects. *Journal of School Leadership, 20*, 184–207.
Metzger, S. A., & Wu, M.-J. (2008). Commercial teacher selection instruments: The validity of selecting teachers through beliefs, attitudes, and values. *Review of Educational Research, 78*, 921–940.
*Mikitovics, A., & Crehan, K. D. (2002). Pre-Professional Skills Test scores as college of education admission criteria. *The Journal of Educational Research, 95*(4), 215–223.
*Mills, S. A. (2007). *An investigation of the predictive validity of a structured teacher-interview instrument (Unpublished doctoral dissertation)*West Lafayette: Purdue University.
Morgesen, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683–729.
Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art – teacher effectiveness and professional learning. *School Effectiveness and School Improvement, 25*, 231–256.
Muijs, D., Reynolds, D., Sammons, P., Kyriakides, L., Creemers, B. P. M., & Teddlie, C. (2018). Assessing individual lessons using a generic teacher observation instrument: How useful is the international system for teacher observation and feedback (ISTOF)? *ZDM, 50*, 395–406.

*Nelson, B. (2013). *Exploring the relationship between the Ventures for Excellence Teacher Style Profile data and teacher performance (Unpublished doctoral dissertation)*St. Charles, MO: Lindenwood University.
*Novotny, M. T. (2009). *The relationship between TeacherInsight scores and professional development and appraisal system domain scores (Unpublished doctoral dissertation)* Denton, TX: University of North Texas.
OECD (2005). *Teachers matter: Attracting, developing and retaining effective teachers.* Paris, France: OECD Publishing. Retrieved from http://www.oecd.org/education/school/attractingdevelopingandretainingeffectiveteachers-finalreportteachersmatter.htm.
OECD (2017). *Education and training.* Paris, France: OECD Publishing. https://stats.oecd.org/Index.aspx?DataSetCode=EAG_TS_STA.
*Oestreich, T. A. (2016). *A study of the effectiveness of the Applitrack Teacherfit hiring selection tool when compared with a teacher's summative evaluation*Unpublished doctoral dissertationBloomington, IN: Indiana University.
Patrick, H., & Mantzicopoulos, P. (2016). Is effective teaching stable? *The Journal of Experimental Education, 84*, 23–47.
Patterson, F., Ashworth, V., Mehra, S., & Falcon, H. (2012). Could situational judgement tests be used for selection into dental foundation training? *British Dental Journal, 213*, 23–26.
Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education, 50*, 36–60.
Paufler, N. A., & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms implications for value-added analyses and interpretations. *American Educational Research Journal, 51*, 328–362.
Ployhart, R. E., & Schneider, B. (2012). The social and organizational context of personnel selection. In N. Schmitt (Ed.). *The Oxford handbook of personnel assessment and selection* (pp. 48–67). Oxford: Oxford University Press.
van Rhee, H., Suurmond, R., & Hak, A. (2015). *User manual for meta-essentials: Workbooks for meta-analyses.* Retrieved from http://hdl.handle.net/1765/78635, Version 1.0.
Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*, 353–387.
Robertson-Kraft, C., & Duckworth, A. (2014). True grit: Trait-level perseverance and passion for long-term goals predicts effectiveness and retention among novice teachers. *Teachers College Record, 116*, 1–25. Retrieved from http://www.tcrecord.org.
*Robison, S. E. (2003). *An exploratory study of the utility of a structured interview in the employment of classroom teachers (Unpublished doctoral dissertation)*West Lafayette: Purdue University.
Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638.
*Ross, S. M. (2013). *Teacher selection in the accountability era of public education: A study of teacher dispositions and student achievement (unpublished doctoral dissertation)* Minneapolis, MN: Capella University.
Ryan, A. M., & Ployhart, R. E. (2014). A century of selection. *Annual Review of Psychology, 65* 20.1-20.25.
Sackett, P. R., Lievens, F., Van Iddekinge, C. H., & Kuncel, N. R. (2017). Individual differences and their measurement: A review of 100 years of research. *Journal of Applied Psychology, 102*, 254–273.
Sahlberg, P. (2015). *Finnish lessons 2.0* (2nd ed.). New York: Teachers College Press.
*Sawyer, G. B. (2005). *A study using the Star Teacher Selection Interview to predict the successful performance of teachers in South Carolina's program of alternative certification for educators (Unpublished doctoral dissertation)*Columbia, SC: University of South Carolina.
Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.
Schwager, I. T., Hülsheger, U. R., Bridgeman, B., & Lang, J. W. (2015). Graduate Student Selection: Graduate record examination, socioeconomic status, and undergraduate grade point average as predictors of study success in a western European University. *International Journal of Selection and Assessment, 23*, 71–79.
Sclafani, S. K. (2015). Singapore chooses teachers carefully. *Phi Delta Kappan, 97*(3), 8–13.
Shultz, M. M., & Zedeck, S. (2012). Admission to law school: New measures. *Educational Psychologist, 47*, 51–65.
Sidik, K., & Jonkman, J. N. (2006). Robust variance estimation for random effects meta-analysis. *Computational Statistics & Data Analysis, 50*, 3681–3701.
Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *The Journal of Economic Perspectives, 24*, 97–118.
Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality, 13*, 389–408.
Strong, M., Gargani, J., & Hacifazlioglu, O. (2011). Do we know a successful teacher when we see one? Experiments in the identification of effective teachers. *Journal of Teacher Education, 62*, 367–382.
Sturman, M. C. (2012). Employee value: Combining utility analysis with strategic human resource management research to yield strong theory. In N. Schmitt (Ed.). *The Oxford handbook of personnel assessment and selection* (pp. 768–792). Oxford: Oxford University Press.
Synar, E., & Maiden, J. (2012). A comprehensive model for estimating the financial impact of teacher turnover. *Journal of Education Finance, 38*, 130–144.
Taylor, N., Mehra, S., Elley, K., Patterson, F., & Cousans, F. (2016). The value of situational judgment tests for assessing non-academic attributes in dental selection. *British Dental Journal, 220*, 565–566.
Thomson, D., Cummings, E., Ferguson, A. K., Moizumi, E. M., Sher, Y., Wang, X., ... Childs, R. A. (2011). A role for research in initial teacher education admissions: A case study from one Canadian university. *Canadian Journal of Educational Administration and Policy, 121*, 1–23.
U.S. Department of Education (2016). *Improving teacher preparation: Building on innovation.* Retrieved from http://www.ed.gov/teacherprep.
UK House of Commons Education Committee (2012). *Great teachers: Attracting, training and retaining the best. Government response to the committee's ninth report of session 2010-2012*London: House of Commons.
Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39*, 111–133.
Villegas, A. M., & Irvine, J. J. (2010). Diversifying the teaching force: An examination of major arguments. *The Urban Review, 42*, 175–192.
Watt, H. M. G., Richardson, P. W., & Wilkins, K. (2014). Profiles of professional engagement and career development aspirations among USA preservice teachers. *International Journal of Educational Research, 65*, 23–40.
Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. New Teacher Project.* New York: The New Teachers Project.
White, B. R., DeAngelis, K. J., & Lichtenberger, E. J. (2013). *The student has become the teacher: Tracking the racial diversity and academic composition of the teacher supply pipeline. Policy research: IERC 2013-3.* Illinois Education Research Council.
*Winters, M. J. (2017). *The relationship between teachers' performance on the Ventures for Excellence and their teaching effectiveness*Unpublished doctoral dissertationPhoenix, AZ: Grand Canyon University.
Wise, A. E., Darling-Hammond, L., & Berry, B. (1987). *Effective teacher selection: From recruitment to retention.* Santa Monica, CA: RAND.
Woolf, K., Potts, H., Stott, J., McManus, C., Williams, A., & Scior, K. (2015). The best choice? *The Psychologist, 28*, 730–735.
Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57–67.
*Young, I. P., & Delli, D. A. (2002). The validity of the Teacher Perceiver Interview for predicting performance of classroom teachers. *Educational Administration Quarterly, 38*, 586–612.
Included in meta-analysis.