Published in *Economics of Education Review 83(August 2021)*

Evidence on the Dimensionality and Reliability of Professional References' Ratings of Teacher Applicants

Dan Goldhaber [a,b], Cyrus Grout [b *], Malcolm Wolff [b], Patrícia Martinková [c, d]

[a] *American Institutes for Research/CALDER, 3876 Bridge Way N, Suite 201, Seattle, WA 98103, United States*

[b] *Center for Education Data and Research, University of Washington, 3876 Bridge Way N, Suite 201, Seattle, WA 98103, United States*

[c] *Faculty of Education, Charles University, Magdalény Rettigové 4, Prague 116 39, Czech Republic*

[d] *Department of Statistical Modelling, Institute of Computer Science of the Czech Academy of Sciences, Pod Vodárenskou věží 2, Prague 182 07, Czech Republic*

[*] *Corresponding author*

This work is supported by the Institute of Education Sciences (grant # R305A170060).

Abstract

There is growing interest in using measures of teacher applicant quality to improve hiring decisions, but the statistical properties of such measures are not well understood. We use unique data on structured ratings solicited from the references of teacher applicants to explore the dimensionality of measures of teacher applicant quality and the inter-rater reliability of the reference ratings. Despite questions about applicants designed to capture multiple dimensions of quality, factor analysis suggests that the reference ratings only capture one underlying dimension. Point estimates of inter-rater reliability range between 0.23 and 0.31 and are significantly lower for novice applicants. It is difficult to judge whether these levels of reliability are high or low in the current context given so little evidence on applicant assessment tools.

Keywords: educational economics; human capital; teacher hiring.

JEL codes: I29, O15

1. Introduction

When hiring teachers, school principals (or other district hiring officials) are certainly selecting teacher applicants on what appear to be multiple dimensions of quality. Principals, for instance, report seeking to hire teachers with good classroom management skills, cultural competence, a strong work ethic, and in-depth subject knowledge (Jacob and Lefgren, 2005; Harris and Sass, 2009; Harris et al., 2010; Giersch and Dong, 2018). Some of these dimensions (or traits) may be associated with readily observable applicant attributes, such as teaching experience and performance on licensure tests. But hiring officials also make judgments about prospective teachers based on subjective assessments of applicant materials that describe hard-to-quantify attributes, such as caring for student well-being, and ability to communicate with and inspire students. This raises the question, to what extent can school systems collect meaningful information about these types of applicant traits?

As described more extensively below, there is a growing interest in systematic measures of teacher applicant quality (Goldhaber et al., 2017; Jacob et al., 2018; Sajjadiani et al., 2018; Bruno and Strunk, 2019) and the type of instruments that school districts can use to rate or pre-screen teacher applicants. Understanding the inter-rater reliability of instruments used to inform teacher hiring is important as there is a direct relationship between the reliability of a measure and the extent to which it will exhibit predictive validity (Martinková et al., 2018). Yet there is little evidence on either the reliability of these measures of applicant quality, or even the degree to which information solicited by school systems is identifying different dimensions of quality ("applicant dimensionality").

Our research focuses on a (likely universal) way that school systems solicit information about teacher applicants: by seeking input from their professional references. The practice of

collecting letters of recommendation from job applicants' references is widespread in the labor market (Aamodt et al., 1993; Salgado, 2001) and as discussed in **Section 2**, in the case of the teacher labor market in particular, there is some evidence that information provided by references is predictive of performance (Goldhaber et al., 2017). Given the ubiquity of the practice of soliciting information from references, understanding the properties of structured reference ratings is of policy interest. Collecting *ratings* from references is a low-cost, easy-to-implement means of information available to hiring officials about applicants, but there is little known about this type of information.[1]

We present evidence from a survey completed by the references of applicants (those who write letters of recommendation for the applicant) to teaching positions in Spokane Public Schools (henceforth, Spokane), a medium-sized urban school district in Washington State. The survey, the development of which is described in **Section 3**, is designed to solicit information about various dimensions of applicant quality. Specifically, references are asked to rate teacher applicants relative to their peers on six competencies thought to be related to effective teaching, to identify which competency is the area of greatest strength and greatest weakness, and to rate each applicant overall.

We find that the distribution of ratings reflects a substantial amount of "cheerleading" – a tendency for references to portray applicants positively, perhaps overly positively – and that the prevalence of cheerleading varies according to rater type (e.g., for principals compared to colleagues). Regarding dimensionality, factor analysis indicates that the reference ratings capture

[1] The collection of applicant ratings from references is distinct from the centralized screening systems studied by Jacob et al. (2018) and Bruno and Strunk (Bruno and Strunk, 2019), which require one-on-one interactions with *district administrators*.

only one underlying dimension of applicant quality. Point estimates of inter-rater reliability range between 0.23 and 0.31 depending on the criteria upon which applicants are being rated; the reliability is significantly higher for experienced applicants relative to novice applicants and for applicants with prior experience in Spokane relative to applicants with out-of-district teaching experience only.

2. Background: Teacher Applicant Information

Through hiring, school districts play a key role in influencing the composition of the teacher workforce. They determine the information applicants are required to provide, the design of screening and interview protocols, and how applicant information is used to inform hiring decisions. For this reason, and as reflected by the literature discussed below, there is a growing interest in the potential for systematic measures of applicant quality to better inform teacher hiring decisions.

Applicant ratings tools are widely used, especially by larger school systems (Metzger and Wu, 2008),[2] but only a few recent studies have examined their ability to predict inservice teacher outcomes.[3] Jacob et al. (2018) studied a multi-stage screening process used by Washington DC Public Schools that included standardized evaluations of applicants based on their taking a written assessment of pedagogical and content knowledge, personal interviews, and

[2] Examples of commercial teacher applicant assessment tools include Gallup's Teacher Insight and Teacher Perceiver tools, the Haberman Foundation's Star Teacher Pre-Screener, and Frontline's series of applicant assessments (see https://www.frontlineeducation.com/blog/applicant-screening-assessments-faqs/, accessed January 29, 2019).

[3] This stands in contrast to the now large body of evidence on the statistical properties of inservice teacher performance measures. For instance, a Bill and Melinda Gates Foundation Study (2012) analyzed the inter-rater reliability of five observation-based teacher evaluation tools, and Hill, Charalambous, and Kraft (2012) conduct a generalizability study of the Mathematical Quality of Instruction, an instrument for measuring mathematics instruction.

teaching auditions. The authors find that composite measures of applicant quality derived from the information collected during the screening process are significantly predictive of future performance as measured by a teacher's IMPACT score – a composite of observational performance evaluations, student progress measures, and (when available) teacher value-added.[4] These findings are generally robust to models that control for selection into the sample.

Bruno and Strunk (2018) examined the link between applicant screening data collected by Los Angeles Unified School District (LAUSD) and outcomes for newly-hired teachers. LAUSD's centralized screening process is used to narrow the pool of applicants eligible for a site-based interview. Rubrics are used to score applicants on a structured interview, sample lesson, written responses to student-related scenarios, professional reference ratings, subject-area preparation and academic background. Here too the authors found that applicants' overall screening performance is significantly predictive of future teacher outcomes, including teacher value-added in English language arts (ELA), observation-based performance, teacher attendance, and the propensity to stay in a school vs switching schools or leaving the district.[5]

Sajjadiani et al. (2018) analyzed a different type of applicant data: detailed work history information provided by applicants to the Minneapolis Public School District (MPSD). They used machine learning techniques to generate measures of work experience relevance, tenure history, and attributions for previous turnover. A Heckman regression approach was adopted to account for the potential bias introduced by sample selection. The authors found strong

[4] When the applicant measures are pooled into an index of predicted performance, the authors find a strong relationship with teacher IMPACT scores: the performance of teachers in the top quartile of predicted performance is 0.71 standard deviations higher than those in the bottom quartile.

[5] The authors do not directly control for selection bias, but report that they do not find evidence that selection is driving results.

connections between their measures of work history and future observational ratings, student evaluations, teacher value-added, and both voluntary and in-voluntary turnover.

Goldhaber, Grout, and Huntington-Klein (2017) found that scores on a job-level applicant screening rubric used by Spokane Public Schools were significantly predictive of teacher outcomes. The screening rubric consisted of scores on ten different criteria, and the authors found that several *individual criteria* had a particularly strong relationship with teacher outcomes: classroom management was strongly predictive of value-added measures of teacher effectiveness in mathematics and reading, and instructional skills, training, and flexibility were strongly predictive of student achievement in mathematics. Scores on several criteria were predictive of retention, including experience, classroom management, flexibility, instructional skills, and interpersonal skills. Part of the ratings process used by Spokane involves assessing information in letters of recommendation written by applicants' professional references and an important implication of the findings of Goldhaber, Grout, and Huntington-Klein (2017) is that applicants' professional references are a useful source of information. This led to the development of the reference ratings tool described in **Section 3** that is the subject of analysis in the current paper.

The studies discussed above assess the predictive validity of measures of applicant quality and teacher outcomes. They do not, however, examine the reliability of these measures – a property that affects the extent to which a measure will exhibit predictive validity. As described by Schmidt and Hunter (1996), the observed correlation between any two measures r_{xy} is attenuated by the reliability of those measures:

$$r_{xy} = r_{x_l y_l}\left(r_{xx} r_{yy}\right)^{1/2}, \tag{1}$$

where r_{xy} is the observed correlation, $r_{x_l y_l}$ is correlation between the true scores of x and y, and r_{xx} and r_{yy} are the respective reliabilities of x and y. Assessing the reliability of a measure is important to understanding how to improve the predictive validity of the measure. For instance, a measure of applicant quality with a low level of inter-rater reliability might be enhanced by providing raters with more training or by increasing the number of raters.

The research that comes closest to the work we present here is Martinková et al. (2018), which examined the inter-rater reliability of applicant ratings from a screening rubric used by school-level hiring officials (typically principals) to identify which applicants to interview in person.[6] Applicants were rated based on information available in their application profiles, including prior experience, training, and letters of recommendation. The authors adopted a mixed-effect model approach to model inter-rater reliability in a hierarchical design, and to test differences in inter-rater reliability. They found that the *within-school* inter-rater reliability of the summative rating was significantly higher for applicants from within the district (0.51) than for those from outside the district (0.42) They also found that the within-school reliability was relatively low for some dimensions of applicant quality – on "cultural competency," for instance, it was only 0.35 for applicants from within the district and 0.33 for applicants from outside the district.[7]

[6] This rubric was also the subject of study in Goldhaber et al. (2017).

[7] There is also some evidence on the properties of hiring rubrics from non-schooling contexts. McCarthy and Goffin (2001), for instance, examine the predictive validity of PR's assessments of applicants to the Canadian Military and Liu et al. (2009) studied applicants to a graduate internship program, but neither of these studies assessed the dimensionality or reliability of the instruments they were studying. See also on the properties of personal reference ratings in the context of applications to graduate school programs (e.g., Oliveri et al., 2017; McCaffrey et al., 2018).

One explanation for the seemingly low level of reliability found by Martinková et al. (2018) is that schooling officials were required to interpret how professional references felt about the teachers for whom they were writing letters of recommendation. As noted above, applicants were assessed, in part, based on the information in their letters of recommendation – which tends to require a certain amount of reading between the lines (e.g., Albakry, 2015). The utility of the information provided by references may be improved if it is collected in the form of a structured survey and generates responses that are easier for hiring officials to interpret.

3. The Application Process and the Collection of Reference Ratings

The first step for individuals wishing to apply for a job in Spokane Public Schools is to create an applicant profile in the online applicant management system. In their profiles, applicants provide information including the following: educational background, qualifications, professional and volunteer experience, personal statements, job preferences, and contact information for at least three references who will provide letters of recommendation. Confidential letters of recommendation are obtained directly from the applicants' PRs, who receive an auto-generated e-mail from the Spokane directing them to an online submission form. That form records the letter writer's name, e-mail address, and relationship to the applicant. References indicate their relationship to applicants by selecting on of the following options: "Principal, Assistant Principal, Principal Assistant, Supervisor, Director"; "University Supervisor"; "Instructional Coach, Department Chair"; "Supervising Teacher during student teacher placement"; "Colleague"; "Other". Having completed a profile, applicants can apply to any number of specific job postings.

To narrow the pool of applicants who will be more closely considered for a position, school principals request that HR provide a reduced list of applicants based on their possessing

certain qualifications selected by the principal, such as having a particular endorsement or type of experience. To determine which applicants are interviewed in person, schools carry out a second stage of screening on the truncated list of applicants: school-level hiring teams (typically including a principal) score each applicant using a district-developed screening rubric that is informed by reviewing information in applicants' profiles, including letters of recommendation. The highest scoring applicants are invited for in-person interviews.

In June 2015, as part of a collaboration with Spokane designed to study and improve teacher hiring practices, we began collecting structured assessments of applicants from their references. Every new reference listed by an applicant receives an email prompting the online submission of letter of recommendation using a provide web link. Following the submission of a letter of recommendation, references are redirected to an online survey where they are asked to rate the applicant relative to his or her peers on a series of criteria (see **Figure 1**). Specifically, the reference is asked the following: "Based on your professional experience, how do you rate this candidate relative to his/her peer group in terms of the following criteria?" For each criterion, the references can rate the candidate as one of the following: "Among the best encountered in my career (top 1%)"; "Outstanding (top 5%)"; "Excellent (top 10%)"; "Very good (well above average)"; "Average"; "Below average"; "No basis for judgement". Four follow-up questions solicit more general assessments from the references:

1. Please select the teaching competency in which the candidate is strongest.
2. If you had to choose, in which competency would you say the applicant is weakest?
3. Overall, how would you rate the candidate?
4. Is there anything else you feel we should know about the applicant? (response optional)

Thank you for taking this additional step to help us better understand the skills and qualifications of applicants to SPS. This short survey shouldn't take more than 5 minutes to complete. Your responses are **confidential** and will **never** be shared with the applicant you are rating.

Based on your professional experience, how do you rate this candidate **relative to her/his peer group** in terms of the following criteria *(hover the cursor over each criterion for further description)*?

Reference name: **TEST**

(Hover over category for description)	Among the best encountered in my career (top 1%)	Outstanding (top 5%)	Excellent (top 10%)	Very Good (well above average)	Average	Below Average	No Basis For Judgement
Challenges Students	○	○	○	○	○	○	○
Classroom Management	○	○	○	○	○	○	○
Working with Diverse Groups of Students	○	○	○	○	○	○	○
Interpersonal Skills / Collegiality	○	○	○	○	○	○	○
Student Engagement	○	○	○	○	○	○	○
Instructional Skills	○	○	○	○	○	○	○

Please select the teaching competency in which the candidate is STRONGEST.

Please Select One ▼

If you had to choose, in which competency would you say the applicant is WEAKEST?

Please Select One ▼

Overall, how would you rate the candidate?

Among the best encountered in my career (top 1%)	Outstanding (top 5%)	Excellent (top 10%)	Very Good (well above average)	Average	Below Average	No Basis For Judgement
○	○	○	○	○	○	○

Is there anything else you feel we should know about the applicant? (response optional)

Submit

Figure 1. Professional Reference Survey Form

Criterion	Description
Challenges Students	• Sets high expectations and holds students accountable
Classroom Management	• Develops routines and procedures to increase learning. • Is effective at maintaining control of the classroom (this may not mean quiet and orderly, but planned and directed) • Students in class treat one another with respect
Working with Diverse Groups of Students	• Is effective at encouraging and relating to students from disadvantaged backgrounds
Interpersonal Skills	• Develops and maintains effective working relationship with colleagues • Contributes to establishing a positive classroom and school environment • Interactions with parents are productive
Student Engagement	• Lessons interest and engage students • Teacher is effective at relating to students
Instructional Skills	• Establishes clear learning objectives and monitors progress • Teacher utilizes multiple approaches to reach different types of students • Ability to adapt curriculum and teaching style to new state and federal requirements

Table 1. Description of Ratings Criteria in References Ratings Survey

The criteria on which applicants are rated consist of teaching competencies with empirically demonstrated links to student achievement and/or other competencies that are of interest to Spokane. These competencies, described in **Table 1**, are: "Classroom Management", "Instructional Skills", "Interpersonal Skills", "Challenges Students", "Student Engagement", and "Working with Diverse Groups of Students". Three of these competencies were demonstrated in previous work to be significantly predictive of teacher value added (Goldhaber et al., 2017): "Classroom Management", "Instructional Skills", and "Interpersonal Skills". Two others, "Student Engagement" and "Challenges Students", are selected on the basis of evidence on the Tripod survey instrument (developed by Ron Ferguson), which measures student perceptions of the classroom instructional environment (Bill & Melinda Gates Foundation, 2010). The last

criterion, “Working with Diverse Groups of Students”, does not have strong evidence linking it to student achievement, but addresses educational equity issues that are of interest to Spokane.

The reference rating survey is designed to be brief, such that a reference can complete it in several minutes. The relative percentile rating method, as well as the questions forcing the reference to identify the competencies in which an applicant is strongest and weakest, are intended to solicit responses exhibiting enough variation across applicants for hiring officials to differentiate between strong and weak applicants (McCarthy and Goffin, 2001).

Since most references probably have positive relationships with their applicants and want to see them do well, it would not be surprising to see applicants described very positively, a pattern henceforth referred to as “cheerleading”. Therefore, we concentrated the ratings categories in the top of the relative percentile distribution (Top 1%, Top 5%, Top 10%, Well Above Average) rather than the bottom (Average, Below Average). This is intended to give references the room to give positive assessments of applicants without always selecting a top rating category. References are also asked two questions that are not subject to cheerleading effects – to select the teaching competencies in which the candidate is strongest and weakest.

Regarding its use by hiring officials, the survey is intended to enhance (rather than replace) other information about the applicant and to allow for a good deal of subjective interpretation. For instance, a hiring official may place more weight on ratings from an applicant’s former principal than on ratings from his or her colleagues. Similarly, some hiring officials may value certain criteria more highly than others depending on the nature of the position they are seeking to fill.

4. Data

From June 2015 to October 2018, we collected 11,527 survey responses (reference ratings) from 3,417 unique applicants and 8,439 unique raters.[8] A plurality of applicants (41%) have three reference ratings, 18% have four, 9% have five, and 4% have six.[9] The majority of raters (85%) rated only one applicant, but a few raters rated 10 or more.

The analytic sample is subject to several sample restrictions, described here. Of the 11,527 survey responses, 314 applicants were rated only once, and 32 applicants were rated 10 or more times. After removing these outliers, which are problematic to the calculation of bootstrapped confidence intervals, we retain 10,842 observations. We also omit 71 ratings where the reference indicated "no basis for judgement" on every criterion and an additional 8 reference ratings where the reference's relationship to the applicant was not recorded. Together these restrictions result in an analytic sample with 10,763 observations, 3,070 unique applicants, 3,601 unique applicant-years, and 8,010 unique references. Since the qualifications and ability of an applicant can be expected to change over time – for instance, an applicant may apply as a novice in 2016 and as an experienced, and more strongly qualified applicant in 2018 – our analysis

[8] As noted above, every new professional reference listed by an applicant during the data collection period received an email prompting the online submission of letter of recommendation and following the submission of a letter, the reference was redirected to our survey form. However, some applicants who applied for jobs during the study period may not have uploaded new letters of recommendation (because they already had three current letters of recommendation, for instance), in which case their references would not have been prompted to complete the survey. The total number of unique applicants during the data collection period from June 2015 to October 2018 was 3,803. And, among professional references redirected to the survey form, we observed a response rate of 95%.

[9] A few survey responses that are included in the study sample are resubmissions (i.e., same applicant and reference); three references made one same-day resubmission, one reference made three same-day resubmissions, three references made same-month resubmissions, and three references made same-year resubmissions. However, there are many applicants who were rated many times without any reference resubmissions.

treats an applicant who received reference ratings in two different years as two different applicants. Henceforth, we use the term "applicant" to refer to an applicant in a specific calendar year.

Below are two descriptive presentations of the survey data. First, we present the relative percentile ratings data. Second, we present data on which competencies are identified as an applicant's strongest and weakest.

4.1 Relative Percentile Ratings Data

Figure 2 shows the distribution of reference ratings for the *Overall* criterion.[10] More than half of applicants are characterized as being "Outstanding (top 5%)" or "Among the best (top 1%)" while fewer than 1% are identified as being "Below average". Given that applicants are likely to request letters of recommendation from individuals with whom they have positive relationships, it is not surprising that our data reflect some amount of cheerleading. While cheerleading is apparent under each type of applicant-reference relationship, we observe substantial variation; references identified as colleagues are the most likely to submit positive ratings while references identified as principals or other administrators are the least likely to do so. For instance, references identified as colleagues awarded a rating of "Among the best (top 1%)" 31% of the time, about twice as often as principals.

[10] Note that ratings criteria for which the reference indicated a rating of "No basis for judgement" are treated as missing values, both in Table 2 and in the analyses described in Section 4. This results in 356 missing values for the student engagement criterion, 457 for instructional skills, 861 for classroom management, 335 for working with diverse students, 18 for interpersonal skills, 524 for challenges students, and 42 for overall. Each sample size is adjusted accordingly according to these missing values in the reliability analysis below.

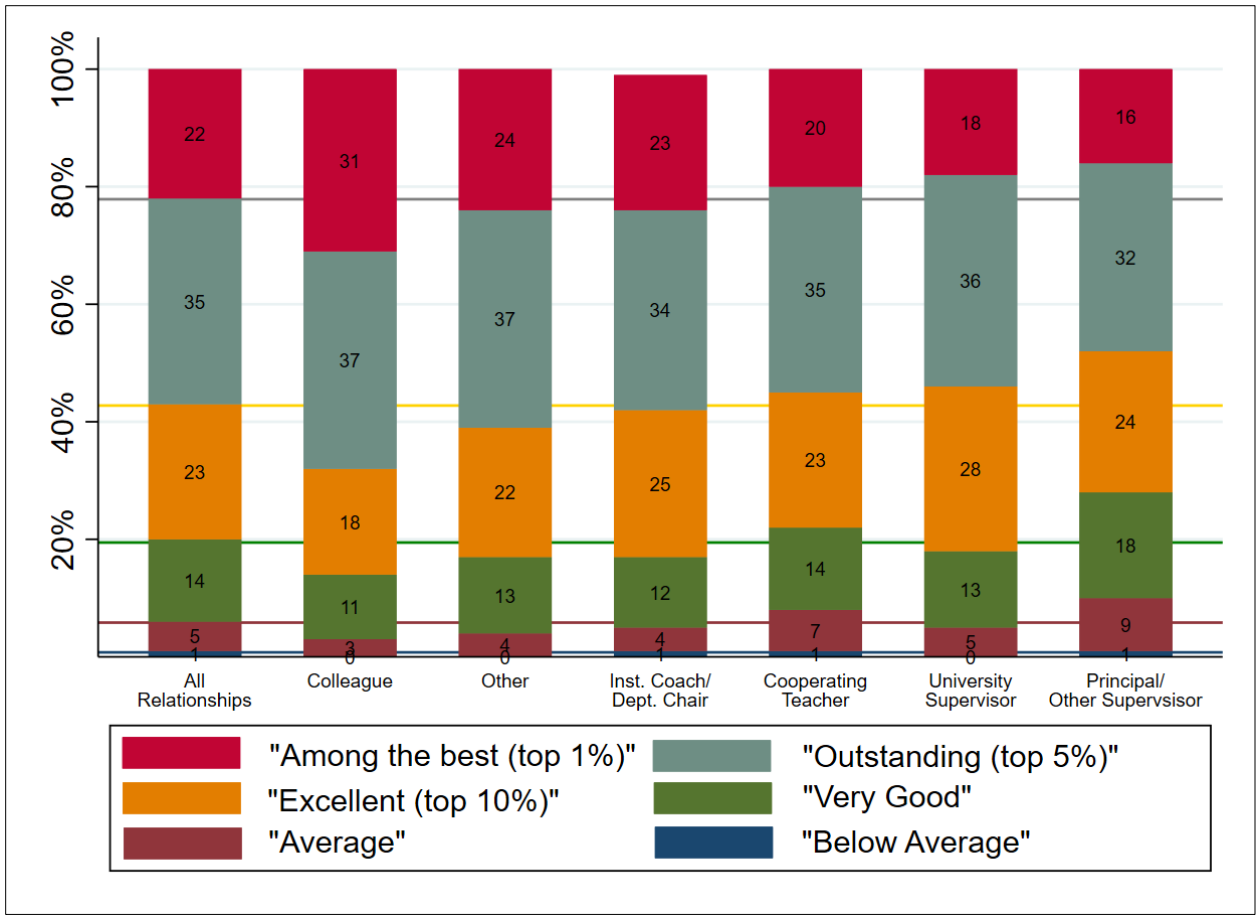

Figure 2: Distribution of Ratings on "Overall" Criterion by Rater Type
Distribution of ratings by applicant-rater relationship type (N = 10,763).

Descriptive statistics for the analytic sample are presented in **Table 2**. We code the relative percentile ratings as follows: "Below average" = 1, "Average" = 2, … , "Among the best (top 1%)" = 6. The average rating for each rater-criterion combination falls between 4 and 5, i.e., between "Excellent (top 10%)" and "Outstanding (top 5%)", and in no case is the average rating for one rater type-criterion combination significantly different from another.[11] The pattern of

[11] Standard deviations range between 1.07 (Colleague-Student Engagement) and 1.31 (Principals/Other Supervisors – Classroom Management). Principals/Other Supervisors' ratings exhibit the greatest variance for every ratings criterion.

principals awarding the lowest scores, and colleagues the highest, persists across each evaluation criterion with one exception – cooperating teachers' ratings of "Classroom Management".[12]

	All Raters	Colleague	Instr. Coach/ Dept. Chair	Cooperating Teacher	Principal/ Other Sup.	University Supervisor	Other
Ratings							
(Average Score on a scale of 1 to 6)							
Challenges	4.52	4.81	4.52	4.35	4.32	4.47	4.67
Management	4.44	4.73	4.50	4.16	4.29	4.35	4.58
Diverse	4.73	4.99	4.77	4.58	4.52	4.69	4.91
Interpersonal	4.73	4.97	4.76	4.71	4.46	4.85	4.87
Engagement	4.65	4.92	4.68	4.53	4.42	4.64	4.83
Instruction	4.53	4.82	4.57	4.39	4.31	4.52	4.66
Overall	4.52	4.83	4.56	4.45	4.25	4.48	4.63
Applicants (Proportions)							
Female	0.69	0.70	0.80	0.68	0.70	0.67	0.66
Internal	0.16	0.17	0.21	0.13	0.19	0.12	0.13
Novice	0.11	0.04	0.06	0.28	0.06	0.30	0.12
Raters (Proportions)							
SPS Employee ("Internal Rater")	0.15	0.13	0.15	0.26	0.18	0.01	0.10
Observations	10,763	2,792	454	1,238	3,598	979	1,702

Table 2. Descriptive Statistics
Descriptive statistics of reference ratings, applicant characteristics, and applicant and rater internal status by reference-applicant relationship. An applicant is "internal" if they have prior teaching experience in Spokane and a rater is "internal" if they are a current employee of Spokane. Ratings are coded as integers between 1 ("Below average") and 6 "Among the best (top 1%)". Ratings criteria for which the reference indicated a rating of "No basis for judgement" are treated as missing values, both in **Table 2** and in the analyses described in Section 4. This results in 356 missing values for "Student Engagement", 457 for "Instructional Skills", 861 for

[12] An applicant's cooperating teacher is the classroom teacher who served as the mentor supervising student teaching – which is why they are a relatively common reference type of among novices.

"Classroom Management", 335 for "Working with Diverse Groups of Students", 18 for "Interpersonal Skills", 524 for "Challenges Students", and 42 for "Overall". Each sample size is adjusted accordingly according to these missing values in the reliability analysis.

Applicants tend to have at least some experience; only 11% report no professional teaching experience, while 16% have teaching experience in the Spokane. Several applicant characteristics are associated with having certain types of references. As one might expect, while novice applicants accounted for 11% of all ratings, only 6% of ratings provided by principals were of novice applicants. Similarly, novice applicants are over-represented among ratings provided by cooperating teachers and university supervisors. Female applicants are over-represented among ratings provided by instructional coaches and department chairs and under-represented among references identified as "Other".

4.2 Strongest and Weakest Competencies

Table 3 displays the frequency with which each competency is identified by raters as an applicant's "Strongest" (Panel A) or "Weakest" (Panel B). Two competencies that stand out are "Challenges Students" and "Classroom Management," which are identified as an applicant's strongest competency with much lower frequency than are other competencies and much more frequently as an applicant's weakest competency. This pattern is particularly strong among raters identified as an applicant's cooperating teachers who identify "Challenges Students" as the strongest competency only 5% of the time and identify "Classroom Management" as the weakest competency 35% of the time. The distributions of competencies identified as strongest and weakest also differ according to relationship type.[13] These differences are likely driven by the fact that different types of raters tend to have known applicants in different types of contexts. For

[13] Chi-squared tests for independence show that the relationship-level distributions are significantly different from one another ($p < 0.01$).

instance, cooperating teachers are more likely to rate "Classroom Management" as an applicant's weakest competency, but as shown in **Table 2**, they are also more likely to be rating novice applicants, who tend to struggle with that skill.

	Panel A – Competency Identified as "Strongest"						
	Obs.	**Challenges**	**Mgmt.**	**Diverse**	**Interpersonal**	**Engagement**	**Instruction**
All Raters	10,763	0.07	0.09	0.22	0.23	0.20	0.19
By Relationship Type							
Colleague	2,792	0.08	0.10	0.25	0.19	0.21	0.16
Inst Coach/Dept Chair	454	0.09	0.11	0.22	0.21	0.18	0.19
Cooperating Teacher	1,238	0.05	0.09	0.21	0.24	0.23	0.18
Principal/Other Sup.	3,598	0.06	0.10	0.24	0.22	0.19	0.20
University Supervisor	979	0.07	0.07	0.17	0.27	0.20	0.22
Other	1,702	0.07	0.06	0.22	0.30	0.22	0.14
	Panel B – Competency Identified as "Weakest"						
	Obs.	**Challenges**	**Mgmt.**	**Diverse**	**Interpersonal**	**Engagement**	**Instruction**
All Raters	10,763	0.26	0.26	0.17	0.15	0.06	0.11
By Relationship Type							
Colleague	2,792	0.26	0.24	0.17	0.17	0.07	0.09
Inst Coach/Dept Chair	454	0.27	0.22	0.17	0.17	0.07	0.10
Cooperating Teacher	1,238	0.24	0.35	0.17	0.09	0.05	0.10
Principal/Other Sup.	3,598	0.26	0.21	0.17	0.18	0.06	0.12
University Supervisor	979	0.24	0.32	0.19	0.09	0.07	0.10
Other	1,702	0.27	0.27	0.14	0.12	0.05	0.13

Table 3: Frequency with which Each Competency Is Rated as "Strongest" or "Weakest"
Proportions should be interpreted by row, such that 8% of raters identified as an applicant's colleague identified "Challenges Students" as the applicant's strongest competency, for example. Rows do not always sum to 100% due to rounding.

To consider how frequently raters agree on what is an applicant's strongest and weakest competencies we restrict the sample to three ratings per applicant and calculate how frequently each of the following scenarios occur: 1) All three raters agree; 2) Two out of three raters agree;

3) No raters agree.[14] The levels of agreement are presented in **Table 4**. For both "Strongest" and "Weakest", we find that all three raters agree for 12% of applicants, two out of three raters agree for 53.5% of applicants, and that each rater identifies a different competency for 34-35% of applicants. These levels of agreement are significantly higher than what would occur at random. If the "Strongest"/"Weakest" competencies identified by each rater were selected at random, the probability of three out of three raters agreeing would be 2.8%, and two out of three raters would agree with a probability of 41.7%.

Please select the teaching competency in which the candidate is STRONGEST.	**Percentage of applicants**
All three raters identify the same competency.	12.0%
Two out of three raters identify the same competency	53.5%
Each rater identifies a different competency	34.5%
If you had to choose, in which competency would you say the applicant is WEAKEST?	**Percentage of applicants**
All three raters identify the same competency.	12.2%
Two out of three raters identify the same competency	53.5%
Each rater identifies a different competency	34.3%
Observations	2,499

Table 4. Level of agreement among raters on applicants' strongest and weakest competencies
Notes: Sample is restricted to exactly 3 ratings per applicant. This restriction excludes 1,102 applicants (with 1,855 ratings) who had fewer than 3 ratings. For applicants with more than 3 ratings, we randomly selected 3 ratings and excluded the remainder, excluding an additional 1,411 ratings from 930 different applicants.

5. Empirical Approach

Our analyses explore the extent to which ratings of teacher applicants by their professional references capture distinct traits of applicant quality, and the inter-rater reliability of the ratings. We describe our approach to these analyses below.

[14] The sample restriction excludes 1,102 applicants (with 1,855 ratings) who had fewer than 3 ratings. For applicants with more than 3 ratings, we randomly selected 3 ratings and excluded the remainder, excluding an additional 1,411 ratings from 930 different applicants.

5.1 Exploratory Factor Analysis of Distinct Traits Captured by Reference Ratings Survey

To examine the extent to which the reference ratings survey measures distinct traits of teacher applicants we perform an exploratory factor analysis. The factor analysis allows us to 1) identify the number of common factors that cause the measures of applicant quality captured by the reference ratings survey to covary, and 2) assess the strength of the relationship between each measure (reference rating) and each identified factor. In the initial exploratory extraction, we do not presume that the ratings data will have a particular number of factors, nor which measures will load onto those factors.

The unadjusted reference ratings data are represented as integers ranging between 1 (*Below average*) and 6 ("Among the best (top 1%)"). Due to the ordinal nature of these data, and the number of value repetitions, we estimate polychoric correlations to perform our factor extraction (see **Appendix A** for further description). Using these correlations, we identify the latent characteristics underlying references' judgements of applicant quality. Formally, the k^{th} professional reference ratings criteria, PRR_k, centered by the mean μ_k, can be described by the equation,

$$PRR_k - \mu_k = l_{k1}F_1 + \cdots + l_{kD}F_D + \varepsilon_k, \qquad (2)$$

for *D* latent factors F_d and mean zero error terms ε_k. Equation (2) is used to identify the loadings l_{kd} that best explain the variance of the reference ratings. As suggested by Costello and Osborne (Costello and Osborne, 2005), we will use a scree test to determine the number of factors to retain.

In addition to examining the dimensionality of the ratings data, the factor loadings derived from the factor analysis are used to generate a summative ratings measure – *PR Factor*.[15] As a robustness check, we also generate a second summative ratings measure ($Theta$) derived from the graded response model (GRM), introduced by Samejima (1969), described in **Appendix B**.

5.2 Inter-Rater Reliability

In the context of our analyses, inter-rater reliability measures the extent to which different references agree about the qualifications of a teacher applicant. Within the framework of generalizability theory (Shavelson and Webb, 1991; Brennan, 2001), each rating is conceived of as a sample from a universe of admissible ratings, which consists of all possible observations that decision makers consider to be acceptable substitutes for the observation in hand.

We analyze inter-rater reliability for the relative percentile ratings of applicants on the six competencies described in Table 1.[16] Due to low percentage of references who rated multiple applicants, we treat raters as nested (and do not include a rater random effect in the model) such that any rater-driven variance is included in the residual error. We also treat the reference rating criteria as fixed and we calculate IRR separately for each criterion using raw reference rating scores as well as for the overall score and the summative ratings described in Section 5.1 – *PR Factor* and *Theta.* This allows for probability-based tests of observed differences in inter-rater reliability across criteria.

---

[15] Specifically, we estimate $\boldsymbol{F}$ from the least-squares regression $\boldsymbol{PRR} - \hat{\boldsymbol{\mu}} = \hat{\boldsymbol{L}}\boldsymbol{F} + \boldsymbol{\epsilon}$, which has analytic solution $\widehat{\boldsymbol{F}} = \left(\hat{\boldsymbol{L}}^T\hat{\boldsymbol{L}}\right)^{-1}\hat{\boldsymbol{L}}^T(\boldsymbol{PRR} - \hat{\boldsymbol{\mu}})$ where $\hat{\boldsymbol{L}}$ is a $K \times D$ matrix of estimated factor loadings from Equation (2) and $\hat{\boldsymbol{\mu}}$ is the empirical mean of the K PR ratings criteria.

[16] In contrast to the relative percentile ratings, the "Strongest"/"Weakest" ratings are unordered categorical responses and cannot be analyzed under the framework presented here.

To estimate inter-rater reliability, we adopt linear mixed-effect regression models (Raudenbush and Bryk, 2002; Goldstein, 2011). As previously discussed, some types of references tend to rate applicants more positively than other types of references (see **Figure 2**). Moreover, Spokane personnel have indicated that hiring officials tend to take these tendencies into account when interpreting the information provided in letters of recommendation. For instance, a rating of "Outstanding (top 5%)" will tend to be interpreted more positively when awarded by an applicant's principal than when awarded by an applicant's colleague, or when awarded by a Spokane employee than when awarded by an outside reference/rater. Therefore, in our primary specification, we account for variation driven by reference type by controlling for the reference-applicant relationship-type fixed effects in the following mixed effect model:

$$PRR_{ij} = \mu + \alpha_1{}'\boldsymbol{r}_{ij} + \alpha_2\, s_j + A_i + \varepsilon_{ij}\,, \qquad (3)$$

where PRR_{ij} is the professional reference rating (criterion, overall or summative) of applicant i by rater j, μ is the overall mean, $\boldsymbol{r}_{ij}$ is a vector of indicators describing a reference's j relationship to the applicant i (e.g., principal, cooperating teacher, or field supervisor), s_j is an indicator that the rater j is an Spokane employee, A_i are applicant-year random effects with variance σ_A^2, and ε_{ij} is a mean zero error term with variance σ_ε^2. We then estimate the contribution of variance from applicants in each group using the variance decomposition model:

$$\sigma_{PRR,A}^2 = \sigma_A^2 + \sigma_\varepsilon^2, \qquad (4)$$

where σ_A^2 represents the systematic error-free variance among scores, having accounted for rater-type fixed effects, and σ_ε^2 represents the random error variance, including any uncaptured variance. Finally, we calculate the inter-rater reliability $IRR \in [0,1]$ of the rater-type-adjusted ratings using the equation (*un*adjusted estimates are available in **Appendix C**):

$$IRR = \frac{\sigma_A^2}{\sigma_A^2+\sigma_\varepsilon^2}\,. \qquad (5)$$

Equation (5) represents the proportion of variance in reference ratings attributable to applicants. At the upper bound, if for each applicant i, every rating of applicant i gives the same score, all variation is explained by differences across applicants and $IRR = 1$. As within-applicant variation increases, the proportion of variation explained by differences across applicants declines and IRR decreases. Hence, IRR measures the level of agreement between raters. To understand whether differences in inter-rater reliability across criteria are statistically significant, we use a parametric bootstrap for mixed models to obtain quantile-based 95% confidence intervals from 1,000 iterations. The parametric bootstrap is implemented using the R statistical software function bootMer( ) of the lme4 package (Bates et al., 2015).

Finally, we compare estimated inter-rater reliability for groups of applicants that may be expected to exhibit different levels of reliability. We make two across-group comparisons. First, we compare applicants with teaching experience in the Spokane ("internal applicants") to applicants with teaching experience outside of the Spokane ("external applicants"). We exclude novice applicants from this comparison to avoid conflating any differences driven by internal vs. external status with those driven by experienced vs. novice status. Raters who have observed an applicant teaching in Spokane may interpret the ratings criteria more consistently than raters of applicants without in-district experience and thus exhibit greater inter-rater reliability. Second, we compare applicants with prior teaching experience to applicants who are novices without any professional experience. We anticipate that ratings of novice applicants will exhibit lower inter-rater reliability because they have less of a track record that references can draw upon to form judgements.

To allow estimated inter-rater reliability to vary by applicant type (e.g., internal vs. external), following Martinkova et al. (2018), we include applicant-type fixed effects and allow

the variance terms of the applicant random effects in **equation (3)** to differ by group by assuming the following mixed-effect model:

$$PRR_{ijg} = \mu + \alpha_1{}'\boldsymbol{r}_{ij} + \alpha_2\, s_j + \alpha_3 p_i + A_{ig} + \varepsilon_{ijg}\,, \qquad (6)$$

Where PRR_{ijg} is the rating (criterion, overall, or composite) of applicant i from group $g\epsilon\{0,1\}$ by rater j, μ is the overall mean, $\boldsymbol{r}_{ij}$ is a vector of indicators describing reference j's relationship to applicant i, s_j is an indicator that the rater j is an employee of Spokane, p_i is an indicator that applicant i belongs to group $g = 1$, A_{ig} are applicant-year random effects for applicants from group g with variance σ^2_{Ag}, and ε_{ij} is a mean zero error term with variance σ^2_{ε}. The decomposition described in **equation (4)** then becomes:

$$\sigma^2_{PRR,Ag} = \sigma^2_{Ag} + \sigma^2_{\varepsilon}, \qquad (7)$$

and the inter-rater reliability for these groups is then calculated as:

$$IRR_g = \frac{\sigma^2_{Ag}}{\sigma^2_{Ag}+\sigma^2_{\varepsilon}}\,. \qquad (8)$$

We use bootstrap procedures to calculate confidence intervals around the point estimates for inter-rater reliability and around the differences in inter-rater reliability across groups in order to understand whether differences in inter-rater reliability between groups are statistically significant.

6. Results

6.1 Factor Analysis of Reference Ratings

The results from the initial factor extraction are presented in **Table 5**. We find that each reference ratings measure loads onto Factor 1 and that the loadings are of similar magnitude

(between 0.87 and 0.96). We also find that the great majority of covariation is driven by Factor 1, as evidenced by Factor 1's large eigenvalue (5.13) and the small eigenvalues of subsequent Factors (see **Figure 3**). In fact, Factor 1 explains 97% of cumulative variation.[17]

Criterion	Unrestricted Model					One-Factor Model
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 1
Challenges	0.94	-0.11	-0.06	0.08	-0.04	0.94
Management	0.93	-0.09	-0.03	-0.09	0.05	0.92
Diverse	0.91	0.16	-0.09	0.03	0.03	0.90
Interpersonal	0.87	0.19	0.08	0.01	-0.01	0.86
Engagement	0.96	-0.01	0.01	-0.08	-0.06	0.96
Instruction	0.94	-0.12	0.09	0.05	0.03	0.95
Cumulative Variation	0.97	0.99	0.99	1.00	1.00	1.00
Eigenvalue	5.13	0.10	0.03	0.02	0.01	5.11

Table 5. Factor Loadings from Initial Factor Extraction.
Note: The left-hand panel reports the solution from an unrestricted model. The right-hand panel reports the solution from a model restricted to one factor.

We use a "scree test" to assess the number of factors underlying covariance in the six reference ratings criteria. As described in Costello and Osborne (2005), a scree test involves plotting the eigenvalues for each sequential factor and looking for a natural break, after which the curve flattens out. We see in **Figure 3** that this break point is located at Factor 2, suggesting that only the Factor 1 be retained for rotation.[18]

[17] As an additional measure of criterion similarity, we conduct linear regression on each ratings criterion including one or multiple other criteria as covariates. Using a single criterion as a covariate, we find that the average regression coefficient is 0.81 across all regressions and ranges from 0.74 (regressing "working with diverse groups of students" on "classroom management") to 0.90 (regressing "instructional skills" on "classroom management"). In regressions with multiple covariates, we find all coefficients are relatively similar, with "working with diverse groups of students" displaying the most substantial deviation.

[18] A critique of the scree test advanced by Courtney (2013), who proposes a series of more technical tests in favor of the scree test, is that it suffers from ambiguity when there is no clear break in the depicted eigenvalues. Such ambiguity is not present in **Figure 3**.

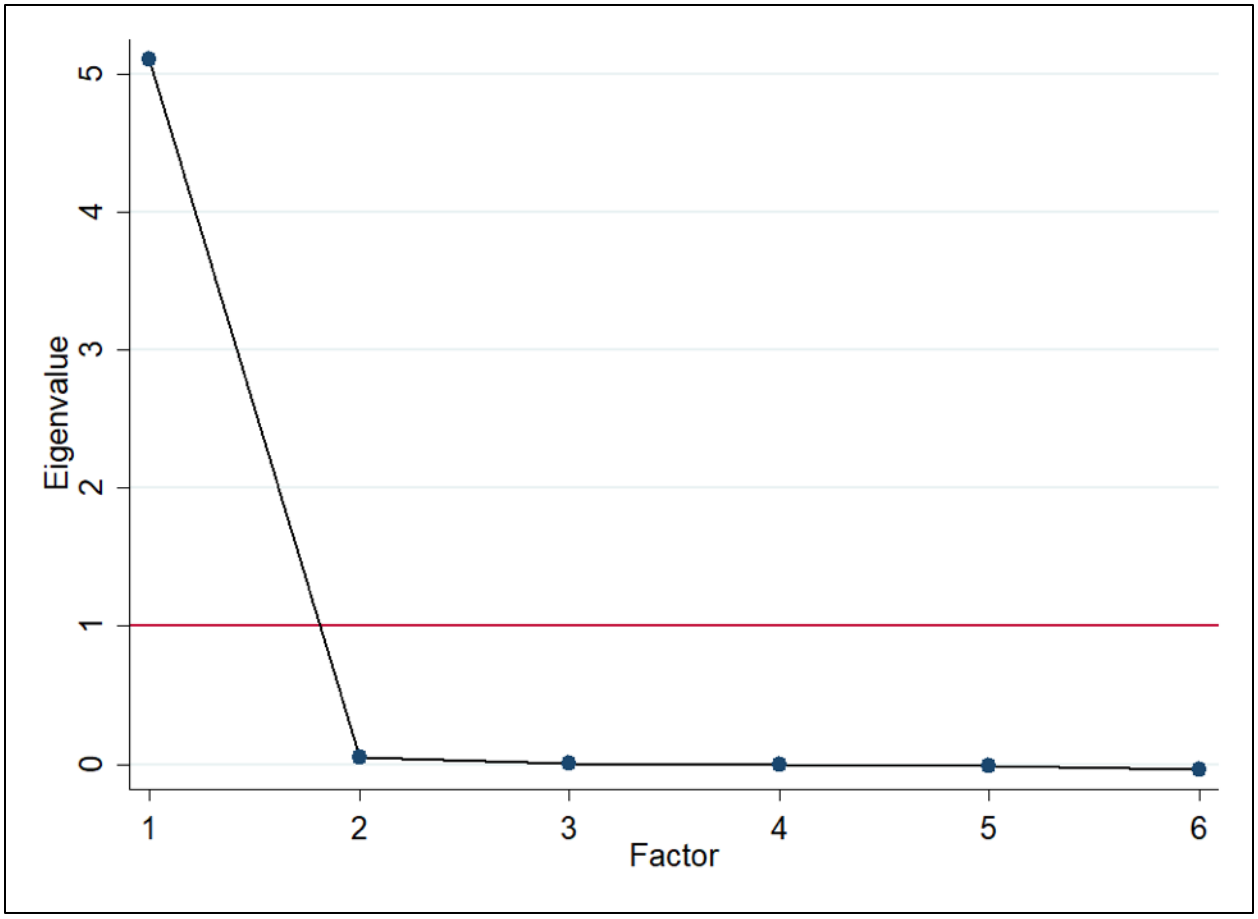

Figure 3. Scree plot of eigenvalues

Note: Scree plot of eigenvalues for six factors generated by the rating matrix including the categories Engagement, Instruction, Management, Diverse, Interpersonal, and Challenges.

The above findings clearly suggest there is just one underlying dimension of applicant quality measured by the reference ratings survey, but it is possible that the dimensionality of the ratings varies by rater or applicant types. We assess this by performing the factor analysis separately for different categories of raters and applicants. Consistent with the findings for the overall sample, we find no evidence that there is more than one dimension for any subsample defined by rater type or applicant type. The factor loadings are also similar across rater and applicant types.

It is also possible that the dimensionality of the ratings is understated due to "halo effects." As described by Oliveri et al. (2017), "Halo effects may arise if an evaluator has a

positive appraisal of the applicant on one trait and then generalizes this positivity to all other traits" (p. 299). As we show in **Table 6**, the correlations across the different dimensions that PRs are asked to rate applicants are quite high. In fact, 23% of the reference ratings rate the applicant at the same level for every criterion. This may raise questions about how seriously some raters took the task of evaluating applicants. But, when we exclude these cases from the factor analysis, the results still strongly indicate the presence of only one factor.

	Challenges	Mgmt.	Diverse	Interpersonal	Engagement	Instruction	Overall	Factor
Challenges	1.00							
Management	0.87	1.00						
Diverse	0.84	0.81	1.00					
Interpersonal	0.79	0.78	0.80	1.00				
Engagement	0.90	0.88	0.85	0.83	1.00			
Instruction	0.91	0.88	0.83	0.81	0.90	1.00		
Overall	0.90	0.89	0.85	0.86	0.92	0.92	1.00	
Factor	0.93	0.91	0.89	0.87	0.94	0.93	0.92	1.00
Theta	0.92	0.90	0.86	0.83	0.93	0.93	0.91	0.98

Table 6: Correlations of Ratings Criteria and Overall Measures
Coefficients are calculated using polychoric correlations on all non-missing criteria from 10,763 ratings and 3,070 applicant clusters.

As described in Section 5.1, we use the factor loadings to generate the summative ratings measure – *PR Factor*. Because we retain only Factor 1, the term $\left(\hat{\boldsymbol{L}}^T\hat{\boldsymbol{L}}\right)^{-1}$ in the analytic solution $\widehat{\boldsymbol{F}} = \left(\hat{\boldsymbol{L}}^T\hat{\boldsymbol{L}}\right)^{-1}\hat{\boldsymbol{L}}^T(\boldsymbol{PRR} - \widehat{\boldsymbol{\mu}})$ reduces to a scalar equal to the sum of squared loadings of the first factor $\sum_k l_{k,1}^2$. This allows us to rescale the PR factor as a weighted average of the original ratings $\sum_k l_{k,1}^2\,\widehat{\boldsymbol{F}} = \hat{\boldsymbol{L}}^T(\boldsymbol{PRR} - \widehat{\boldsymbol{\mu}})$, with weights given by the factor loadings reported in the right-hand panel of **Table 5** , to retain a score range consistent with the original ratings. $PR\ Factor$ ranges between 1.00 and 6.00, has a mean of 4.26 (between "Excellent (top 10%)" and "Outstanding (top 5%)") and standard deviation of 0.998. $PR\ Factor$ is strongly correlated with reference ratings for the criterion "Overall" ($\rho = 0.93$). We also calculate a GRM

linearized transformation of the reference ratings ($Theta$), which is also strongly correlated with the "Overall" criterion ($\rho = 0.93$) and with $PR\ Factor$ ($\rho = 0.98$).

6.2 Inter-Rater Reliability

Results from the estimation of **equation (3)** for each rating criterion and our two composite measures, $PR\ Factor$ and $Theta$, are presented in **Table 7**. For each criterion including "Overall" and for the two composite measures, we find that the type of applicant-reference relationship is a significant source of variation in professional reference ratings. In each case, colleagues rate applicants significantly higher than other types of references. Principals tend to rate applicants lower than other types of raters. For instance, they rate applicants between 15% and 43% of a standard deviation lower than other raters on the summative measure $Theta$ (column (9)). We also find that internal raters tend to rate applicants less positively, though the difference is not always statistically significant. As noted above, because hiring officials are likely to take rater type into consideration when interpreting the ratings of applicants, the inter-rater reliability point estimates presented below are adjusted for these rater-type sources of variation (i.e., they are included as fixed effects r_{ij} and s_j in **equations (3)** and **(6)**). Estimates of inter-rater reliability ***unadjusted*** by rater type are presented in **Table C2** in **Appendix C**.

	(1) Challenges	(2) Management	(3) Diverse	(4) Interpersonal	(5) Engagement	(6) Instruction	(7) Overall	(8) Factor	(9) Theta
Relationship Type									
Colleague	0.506***	0.457***	0.488***	0.536***	0.504***	0.531***	0.605***	0.454***	0.435***
	(0.029)	(0.031)	(0.028)	(0.028)	(0.029)	(0.029)	(0.028)	(0.023)	(0.022)
Instructional Coach/Dept. Chair	0.170***	0.205***	0.225***	0.276***	0.225***	0.237***	0.282***	0.202***	0.175***
	(0.056)	(0.059)	(0.054)	(0.055)	(0.055)	(0.055)	(0.055)	(0.044)	(0.043)
Cooperating Tchr.	0.146***	0.009	0.147***	0.314***	0.212 ***	0.227***	0.311***	0.167***	0.153***
	(0.038)	(0.040)	(0.037)	(0.037)	(0.037)	(0.037)	(0.039)	(0.031)	(0.029)
Principal	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)
Univ. Supervisor	0.213***	0.144***	0.226***	0.432***	0.265***	0.292***	0.279***	0.235***	0.202***
	(0.042)	(0.045)	(0.041)	(0.041)	(0.041)	(0.042)	(0.042)	(0.034)	(0.032)
Other	0.403***	0.357***	0.429***	0.446***	0.442***	0.433***	0.450***	0.404***	0.335***
	(0.037)	(0.040)	(0.035)	(0.034)	(0.036)	(0.036)	(0.035)	(0.028)	(0.027)
Internal Rater	-0.064**	-0.052	-0.005	0.002	-0.080***	-0.075***	-0.083***	-0.040	-0.042
	(0.036)	(0.039)	(0.034)	(0.035)	(0.036)	(0.036)	(0.035)	(0.028)	(0.028)
Intercept	4.307***	4.271***	4.501***	4.444***	4.419***	4.279***	4.236***	3.806***	-0.220***
	(0.022)	(0.022)	(0.021)	(0.022)	(0.022)	(0.022)	(0.024)	(0.018)	(0.019)
Applicant variance	0.3374	0.4364	0.2683	0.3436	0.3463	0.3553	0.383	0.2526	0.2353
Residual variance	1.0358	1.0937	1.0089	1.0138	1.0096	0.9956	0.9656	0.6064	0.6239
Applicant clusters	3,550	3,502	3,573	3,601	3,569	3,564	3,600	3,601	3,601
Observations	10,239	9,902	10,428	10,745	10,407	10,306	10,763	10,763	10,763

Table 7: Mixed effect models with rater-type fixed effects

Each column is a separate regression. The reference category for relationship type is Principal. Ratings are coded as integers between 1 (*Below average*) and 6 "Among the best (top 1%)". The mean and standard deviation of the summative measures $Factor$ are 4.23 and 1.00 respectively. The summative measure $Theta$ is standardized $\sim N(0,1)$. Ratings criteria for which the reference indicated a rating of "No basis for judgement" are treated as missing values. All regressions include applicant random effects. Standard errors shown in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Figure 4 presents the estimated inter-rater reliability for each rating criterion including "overall" rating, and the two summative measures, *PR Factor* and *Theta*. Point estimates range between 0.26 and 0.31 and, generally fall within the margin of error of one another. The exception is the criterion "Working with Diverse Groups of Students," which has a significantly lower point estimate of 0.23. This finding is consistent with Martinková et al. (2018) who found that the inter-rater reliability of a similar criterion called "Cultural Competency" was lower than for other criteria. It may be that the relatively low reliability of the criterion "working with diverse groups of students" is due to a general difficulty that educators have in assessing this type of attribute.

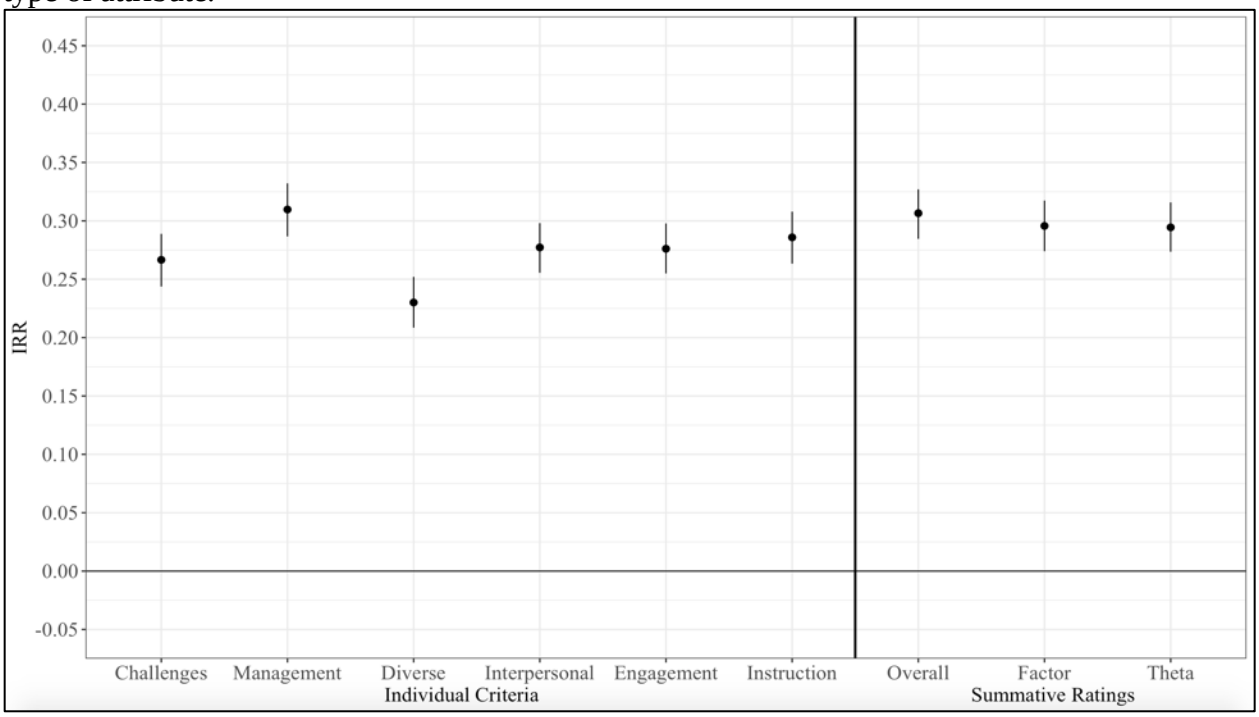

Figure 4: Inter-Rater Reliability by Rating Category
Point estimates of IRR across rating category, "overall" rating, the generated PR Factor, and *Theta*, using 3,601 applicant-years across 10,763 ratings. Confidence intervals are generated using parametric bootstrap with 1,000 replications.

In **Figures 5** and **6** (also see Supplementary Tables **C2** and **C3** in **Appendix C)**, we consider whether ratings for different types of applicants exhibit different levels of inter-rater reliability. First, we compare the inter-rater reliability of reference ratings for internal applicants who report prior experience teaching in Spokane to that of external applicants who report teaching experience outside of Spokane in **Figure 5**. Inter-rater reliability is consistently higher for internal applicants. In each case, the 95% confidence interval around the difference – represented by the black series of bars – is above zero. The largest difference is for the

"Instructional skills" criterion (0.12) and the smallest is for the "Interpersonal skills" criterion (0.05). Again, these findings are quite consistent with Martinková et al. (2018) who found that reliability of the summative rating of internal applicants was significantly higher than that of external applicants (though differences for specific evaluation criteria were not statistically significant).

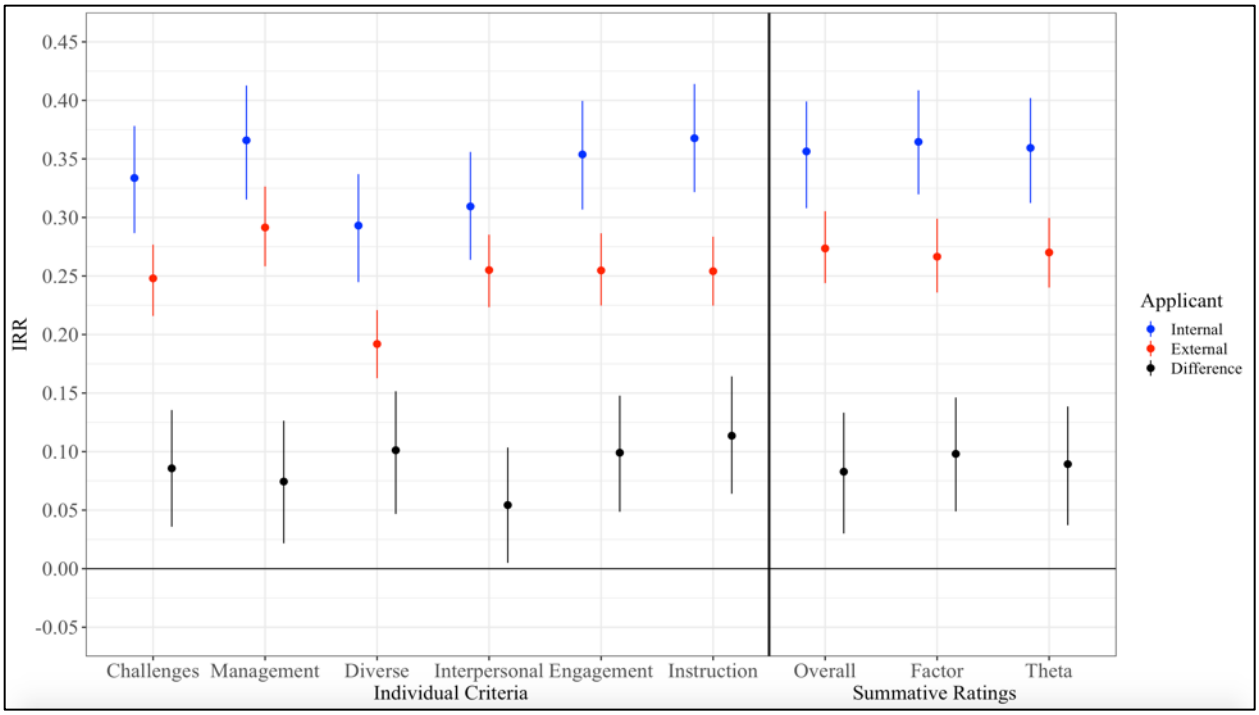

Figure 5: Inter-Rater Reliability by Rating Category and Applicant Type (Internal/External) Point estimates of IRR by applicant internal/external status across rating category, including "overall", the generated PR Factor, and *Theta*, using 3,021 applicant-years across 8,939 ratings (15% of which are internal). Confidence intervals are generated using parametric bootstrap with 1,000 replications.

Second, we compare the inter-rater reliability of ratings for applicants who have prior teaching experience to that of applicants who are novices without any professional experience in **Figure 6**. We find that inter-rater reliability is consistently higher for experienced applicants than for novice applicants and that in some criteria ("Instructional skills", "Classroom management", "Interpersonal skills", "Challenges students") as well as in the two composite measures, the

difference in IRR between novices and experienced applicants is statistically significant. The largest difference is for the "Classroom management" criterion (0.11).

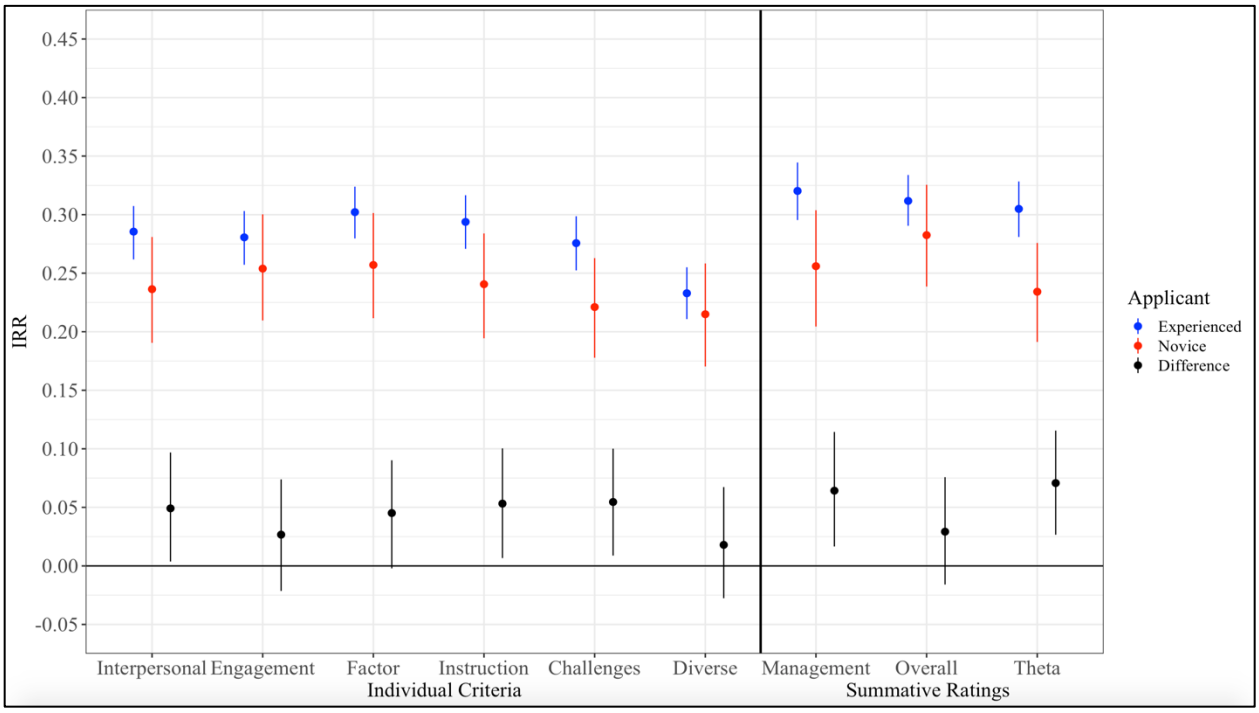

Figure 6: Inter-Rater Reliability by Rating Category and Applicant Type (Novice/Exp)
Notes: Point estimates of IRR by applicant experienced/novice status across rating category, including "overall", the generated PR Factor, and *Theta*, using 3,601 applicant-years across 10,763 ratings (11% of which are novice). Confidence intervals are generated using parametric bootstrap with 1,000 replications.

Given the differences found between different types of applicants, it is natural to ask whether different types of references are more reliable raters. When we restrict the sample to ratings from principals (which requires excluding applicants *without* two or more ratings from Principals), we find estimates of inter-rater reliability for the summative measures $Factor$ and $Theta$ slightly above 0.45 – substantially higher than the estimates in the range of 0.29 and 0.32 for the overall sample.[19] It is possible that principals are in fact more reliable raters, but, as noted

[19] Sample sizes become sparse when restricting the sample to applicants with two or more ratings from references of other relationship-types resulting in very imprecise estimates.

above, this finding may be also driven by the sample construction – the types of applicants that have two or more principals as references may be easier to reliability rate because they have more experience, giving principals more information upon which to base their ratings.

In interpreting the above results, it is important to note that they relate to the "single-rater" reliability of the reference ratings. One way to improve the reliability of a rating is to use the average rating from multiple raters. Following Martinková et al. (2018), we can use the estimated variances from our primary results to consider the "multiple-rater reliability" of an average rating taken from multiple raters. Using the Spearman-Brown Prophecy Formula, when the average rating from m references is used, the variance of the average rating decomposes to

$$\sigma^2_{PRR} = \sigma^2_A + \sigma^2_\varepsilon/m, \qquad (9)$$

and the (multiple-rater) inter-rater reliability is accordingly,

$$IRR = \sigma^2_A/(\sigma^2_A + \sigma^2_\varepsilon/m). \qquad (10)$$

When this formula is applied to the estimated variances obtained from equations (3) and (4), and $m = 3$ raters, we obtain estimates of inter-rater reliability for the individual ratings criteria in the range of 0.47 ("Working with Diverse Groups of Students") to 0.57 ("Classroom Management"), substantially higher than the single-rater reliability estimates presented in **Figure 4**.

7. Discussion and Conclusions

Together, the results presented in Section 6 shed light on the properties of ratings of teacher applicants by their professional references. To our knowledge, this is the first evidence on the properties of applicant ratings in the context of a teacher hiring instrument. We are unsure how widely such instruments are used in the context of job applicant screening but requiring letters of recommendation from professional references is quite common. Given this, and the fact that professional references play a role in the high-stakes decision over whether to hire a job

applicant, understanding the extent to which letter writers can differentiate applicant attributes and/or agree about applicant quality are fundamental issues.

Regarding dimensionality, the finding that only one factor significantly influences the measures captured by the reference ratings survey reflects several possibilities. The first is that there truly is only one underlying trait of applicant quality. This conflicts with previous research on the relationship between teacher applicant information and teacher (and student) outcomes, which suggests multiple dimensions of quality (Rockoff et al., 2011; Jacob et al., 2018; Sajjadiani et al., 2018; Bruno and Strunk, 2019).

If, as seems likely, there are in fact multiple underlying dimensions of applicant quality, they may simply be difficult to identify based on our rating instrument or, more generally, during the hiring process. Regardless, this seems problematic in the case of teacher hiring. There is a growing emphasis on hiring teachers with an ability to connect with a diverse range of students (National Academies of Sciences, Engineering, and Medicine, 2020) and evidence that teacher effectiveness is multidimensional (Kraft, 2019). It is possible that refining the rating instrument would increase its dimensionality, or that information about teacher applicants needs to be derived from other types of assessments, such as sample teaching lessons (e.g., Jacob et al., 2018).

Our analysis of inter-rater reliability finds single-rater reliability estimates that are in the range of 0.23 to 0.31 for individual rating criteria. While the magnitudes of these estimates are well below what is considered to be desirable for high-stakes decisions (Cicchetti, 1994; Hill et al., 2012), it is difficult to judge whether these levels of reliability are high or low in the current context given that there is so little evidence on the reliability of comparable or alternative applicant assessment tools. Cicchetti (1994), for instance, provides the following characterization

of inter-rater reliability for psychological assessment tools: values below 0.40, between 0.40 and 0.59, between 0.60 and 0.74, and above 0.75 are indicative of poor, fair, good, and excellent reliability, respectively. However, different types of tests have been found to exhibit different levels of inter-rater reliability. Lee (2012), for instance, cites single-rater reliability levels for peer reviewed grant proposals in the range of 0.19 to 0.37, and argues that variance in reviewer ratings can be accounted for by normatively appropriate disagreements such as individual differences in areas of expertise, scientific interests, and value systems. Similarly, Erosheva et al. (2021) report values of 0.31 and 0.37 in grant peer review and further discuss statistical issues which may cause low levels of inter-rater reliability. And Rust and Golombok (Rust and Golombok, 2009) note that different types of psychometric tests are subject to different norms for what is an acceptable level of reliability: > 0.9 for intelligence tests, >0.7 for personality tests, ~0.6 for essay marking, and ~0.2 for Rorschach inkblot tests.[20]

As noted above, the research that comes closest to the work we present here is Martinková et al. (2018), which examined the inter-rater reliability of applicant ratings from a screening rubric used by school-level hiring officials. They observed higher levels of inter-rater reliability (in the range of 0.33 to 0.51 depending on the measurement criterion and the internal/external status of the applicant)), but it is important to note that their modeling approach estimated *within*-school inter-rater reliability. Estimates of *across*-school inter-rater reliability, presented in a working paper version of the paper (Martinková and Goldhaber, 2015), were substantially lower. The reference ratings we study in this paper are not associated with any

[20] For a more general overview of the various issues that arise with testing, see American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014).

particular school or job posting. Another important difference between our analysis and that of Martinková et al. (2018) is that the we study different types of raters. The school-level hiring officials studied by Martinková et al. (2018) would have received training on the use of the ratings rubric and would have been likely to hold an understanding of the ratings criteria in common.

In contrast, there are several potential limitations to inter-rater reliability associated with the type of raters we study – professional references. First, they are not positioned to receive training on how to rate applicants. And while the language used to define our ratings criteria is consistent with that used in the context of teacher performance evaluations in the Washington State, it is difficult to know whether raters are interpreting the criteria as intended.[21] Second, raters are likely to have known a particular applicant under different circumstances or during different periods of time (e.g., as their university supervisor versus as a fellow teacher), meaning that references are often forming judgements about the applicant using different sets of information.[22] Finally, we have considered the possibility that some raters are not particularly attentive to the survey and that this is a source of low reliability. However, when we exclude raters whose ratings on the six criteria are logically inconsistent with their identification of the competency in which the applicant is strongest or weakest, we find estimates of inter-rater reliability that are similar to estimates for the overall sample.

[21] Here, our finding of substantially higher levels of reliability when the sample is restricted to references identified as the applicant's *Principal/Other Supervisor* is of interest. In Washington State, conducting teacher performance evaluations is an important part of a principal's job and as a group, they are more likely to have a consistent understanding of the ratings criteria than other types of raters.

[22] Some references will naturally have more knowledge about the applicant than other references. When we exclude survey responses where the rater indicated "No basis for judgement" for one or more criteria, we find estimates of IRR that are slightly higher than for the overall sample.

It is important to recognize that the predictive validity of the ratings of teacher applicants (the extent to which they predict outcomes of inservice teachers) is limited by their reliability (Hill et al., 2012). That said, a lower level of inter-rater reliability may be acceptable in the context of professional reference ratings (as opposed to performance evaluations, for instance) because they constitute one piece of information used to inform a high stakes decision but are not determinative of that decision. In the context of ratings collected from professional references, relatively low levels of reliability are also mitigated by the fact that applicants are typically asked to provide multiple references. As shown in Section 6.2, estimates of inter-rater reliability for the individual criteria range between 0.47 and 0.57 when we consider the average rating taken from three raters rather than single-rater reliability.

Given the evidence on the importance of teacher quality for student achievement, we should further explore the properties of teacher applicant assessment mechanisms and the extent to which various means of judging teacher applicants are linked to the future performance of teachers. Our analysis of inter-rater reliability identified some subgroups where inter-rater reliability is lower – for novice applicants versus experienced ones, and for applicants with external experience versus those with within-district experience. Future work might include efforts to increase inter-rater reliability among these groups. Finally, the finding of only one dimension underlying the survey responses is valuable. It suggests the current practice is wasteful and suggests two possible directions for improvement. The current set of questions could be pared down without losing information and different questions could be developed to try to capture other dimensions of applicant quality.

Acknowledgements

This work is supported by the Institute of Education Sciences (grant # R305A170060) and the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), which is funded by a consortium of foundations. PM was partly supported by Czech Science Foundation Grant 21-03658S. For more information about CALDER funders, see www.caldercenter.org/about-calder. We appreciate comments on an earlier draft by Wai-Ying Chow, Lauren Dachille, and Aaron Sojourner. All opinions expressed in this paper are those of the authors and do not necessarily reflect the views of our funders or the institutions to which the authors are affiliated.

References

Aamodt, M. G., Bryan, D. A., and Whitcomb, A. J. (1993). Predicting Performance with Letters of Recommendation. *Public Personnel Management*, *22*(1), 81–90. doi:10.1177/009102609302200106

Albakry, M. (2015). Telling by omission: Hedging and negative evaluation in academic recommendation letters. In V. Cortes & E. Csomay (Eds.), *Corpus-based research in applied linguistics: Studies in honor of Doug Biber (Vol. 66)* (pp. 79–98). John Benjamins Publishing Company. doi:https://doi.org/10.1075/scl.66

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington D.C.: American Educational Research Association.

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Bill & Melinda Gates Foundation. (2010). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. Seattle, WA. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

Bill & Melinda Gates Foundation. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*. *MET Project Research Paper*. Seattle, WA. Retrieved from http://metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer.

Bruno, P., and Strunk, K. O. (2018). *Making the Cut: The Effectiveness of Teacher Screening and Hiring in the Los Angeles Unified School District* (No. 184). Washington D.C.

Bruno, P., and Strunk, K. O. (2019). Making the Cut: The Effectiveness of Teacher Screening and Hiring in the Los Angeles Unified School District. *Educational Evaluation and Policy Analysis*, *41*(4), 426–460. doi:10.3102/0162373719865561

Cicchetti, D. V. (1994). Guidelines , Criteria , and Rules of Thumb for Evaluating Normed and. *Psychological Assessment*, *6*(4), 284–290. doi:10.1037/1040-3590.6.4.284

Costello, A. B., and Osborne, J. W. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research, & Evaluation*, *10*(7), 1–9. Retrieved from https://methods.sagepub.com/base/download/BookChapter/best-practices-in-quantitative-methods/d8.xml

Courtney, M. G. R. (2013). Determining the Number of Factors to Retain in EFA: Using the SPSS R-Menu v2.0 to Make More Judicious Estimations. Practical Assessment, Research & Evaluation, 18(8). *Practical Assessment, Research & Evaluation*, *18*(8), 1–14.

Erosheva, E. A., Martinková, P., and Lee, C. J. (2021). When zero may not be zero: A cautionary note on the use of inter- - rater reliability in evaluating grant peer review. *Journal of the Royal Statistical Society Series A*, 1–16. doi:10.1111/rssa.12681

Giersch, J., and Dong, C. (2018). Principals' preferences when hiring teachers: a conjoint experiment. *Journal of Educational Administration*, *56*(4), 429–444. doi:https://doi.org/10.1108/JEA-06-2017-0074

Goldhaber, D., Grout, C., and Huntington-Klein, N. (2017). Screen Twice, Cut Once: Assessing the Predictive Validity of Teacher Selection Tools. *Education Finance and Policy*, *12*(2), 197–223. doi:doi:10.1162/EDFP_a_00200

Goldstein, H. (2011). *Multilevel Statistical Models* (4th ed.). Bristol, UK: Wiley.

Harris, D. N., Rutledge, S. A., Ingle, W. K., and Thompson, C. C. (2010). Mix and Match : What Principals Really Look for When Hiring Teachers. *Education Finance and Policy*, *5*(2), 228–246.

Harris, D. N., and Sass, T. R. (2009). *What Makes for a Good Teacher and Who Can Tell?* Retrieved from http://calderprod.urban.org/upload/CALDER-Working-Paper-30_FINAL.pdf

Hill, H. C., Charalambous, C. Y., and Kraft, M. A. (2012). When rater reliability is not enough: teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64. doi:10.3102/0013189X12437203

Jacob, B. A., and Lefgren, L. (2005). *Principals as Agents: Subjective Performance Measurement in Education* (No. 11463). National Bureau for Economic Research.

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B., and Rosen, R. (2018). Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools. *Journal of Public Economics*, *166*, 81–97. doi:https://doi.org/10.1016/j.jpubeco.2018.08.011

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*(1), 1–36. doi:10.3368/JHR.54.1.0916.8265R3

Lee, C. J. (2012). A Kuhnian critique of psychometric research on peer review. *Philosophy of Science*, *79*(5), 859–870. doi:10.1086/667841

Liu, O. L., Minsky, J., Ling, G., and Kyllonen, P. (2009). Using the Standardized Letters of Recommendation in Selection: Results From a Multidimensional Rasch Model. *Educational and Psychological Measurement*, *69*(3), 475–492. doi:10.1177/0013164408322031

Martinková, P., and Goldhaber, D. (2015). *Improving Teacher Selection: The Effect of Inter-*

Rater in the Screening Process (No. 2015–7). Seattle, WA. Retrieved from http://cedr.us/papers/working/CEDR WP 2015-7.pdf

Martinková, P., Goldhaber, D., and Erosheva, E. (2018). Disparities in ratings of internal and external applicants : A case for model-based inter-rater reliability. *PLoS ONE*, *13*(10), 1–17. doi:10.1371/journal.pone.0203002

McCaffrey, D. F., Oliveri, M. E., and Holtzman, S. (2018). A Generalizability Theory Study to Examine Sources of Score Variance in Third-Party Evaluations Used in Decision-Making for Graduate School Admissions. *ETS Research Report Series*, *2018*(1). doi:10.1002/ets2.12225

McCarthy, J. M., and Goffin, R. D. (2001). Improving the Validity of Letters of Recommendation: An Investigation of Three Standardized Reference Forms. *Military Psychology*, *13*(4), 199–222. doi:10.1207/S15327876MP1304_2

Metzger, S. A., and Wu, M.-J. (2008). Commercial Teacher Selection Instruments: The Validity of Selecting Teachers Through Beliefs, Attitudes, and Values. *Review of Educational Research*, *78*(4), 921–940. doi:10.3102/0034654308323035

National Academies of Sciences Engineering and Medicine. (2020). *Addressing Changing Expectations for K-12 Teachers in the United States: Policies, Preservice Programs, and Professional Development*. Washington, D.C.

Oliveri, M., McCaffrey, D., Ezzo, C., and Holtzman, S. (2017). A Multilevel Factor Analysis of Third-Party Evaluations of Noncognitive Constructs Used in Admissions Decision Making. *Applied Measurement in Education*, *30*(4), 297–313. doi:10.1080/08957347.2017.1353989

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Rockoff, J. E., Jacob, B. A., Kane, T. J., and Staiger, D. O. (2011). Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy*, *6*(1), 43–74. Retrieved from http://www.mitpressjournals.org.offcampus.lib.washington.edu/doi/pdf/10.1162/EDFP_a_00022

Rust, J., and Golombok, S. (2009). *Modern psychometrics: The science of psychological assessment* (3rd ed.). Routledge/Taylor & Francis Group. Retrieved from https://psycnet.apa.org/record/2008-09955-000

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., and Mykerezi, E. (2018). *Machine Learning and Applicant Work History*.

Salgado, J. F. (2001). Personnel Selection Methods. In I. T. Robertson & C. L. Cooper (Eds.), *Personnel Psychology and Human Resource Management: A Reader for Students and Practitioners* (pp. 1–54). Manchester, UK: John Wiley & Sons, LTD.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2), 100. Retrieved from https://psycnet.apa.org/record/1972-04809-001

Schmidt, F. L., and Hunter, J. E. (1996). Measurement Error in Psychological Research: Lessons From 26 Research Scenarios. *Psychological Methods*, *I*(2), 199–223.

Shavelson, R. J., and Webb, N. M. (1991). *Generalizability Theory: A Primer* (1st ed.). Newbury Park, CA: Sage Publications, Inc.

Appendix A. Polychoric Correlation

The most commonly used measures of correlation are the Pearson correlation, Spearman correlation and Kendall's Tau, each of which has shortcomings in the context of our data, which is discrete rather than continuous. Pearson correlation requires multivariate normality and hence continuous data, and its use on ordinal data correlation leads to an underestimate of the degree of association between observed values and hence a decrease in factor weights when conducting factor analysis, leading to an underestimate of relative importance when assigning factor weights (Holgado-Tello et al., 2010). The Spearman correlation and Kendall's Tau have been shown to have increased bias and squared error relative to polychoric correlation (Babakus & Ferguson, 1988).

The polychoric correlation is defined as follows: Let U and V be discrete random variables that take on m_U and m_V values, respectively. Polychoric correlation assumes that there exist two underlying latent variables X and Y such that

$$U = i \leftrightarrow \tau_{i-1} \leq X < \tau_i \quad i = 1, 2, \dots, m_U,$$

$$V = j \leftrightarrow \xi_{j-1} \leq Y < \xi_j \quad j = 1, 2, \dots, m_V,$$

where

$$-\infty = \tau_1 < \tau_2 < \cdots < \tau_{m_U} = \infty,$$

$$-\infty = \xi_1 < \xi_2 < \cdots < \xi_{m_V} = \infty,$$

are thresholds, and $\sigma_X^{-1}(X - \mu_X),\ \sigma_Y^{-1}(Y - \mu_Y) \sim N(0, 1)$. The polychoric correlation estimates the unique correlation $\hat{\rho}$ between U and V which minimizes the distance to the theoretical correlation ρ^* between X and Y.

Appendix B. Graded Response Modeling

The GRM was introduced by Samejima (1969, 1972, 1995) to handle ordered categories, such as letter grades, or subjective responses, such as those solicited by Likert scales. The cumulative category response function (CCRF) is given by

$$P_{ik}^{*}(\theta) = P(Y_i \geq k \mid a_i, \boldsymbol{b_i}; \theta) = \frac{\exp\left(a_i(\theta - b_{ik})\right)}{1+\exp\left(a_i(\theta - b_{ik})\right)} \qquad (6),$$

where $P_{ik}^{*}(\theta)$ is the probability of an examinee with proficiency θ scoring at least k on item i. This is a a_i is the discrimination parameter of item i, and $\boldsymbol{b}_i$ is a vector of difficulty parameters of item i. Then the probability of each score is

$$P_{ik}(\theta) = P(Y_i = k \mid a_i, \boldsymbol{b_i};\ \theta) = P_{ik}^{*}(\theta) - P_{ik+1}^{*}(\theta). \qquad (7)$$

Letting $\boldsymbol{B} = (a_1, \ldots, a_I, \boldsymbol{b_1}, \ldots, \boldsymbol{b_I})$, the likelihood is computed by integrating out the latent variable from the joint density:

$$L(\boldsymbol{B}) = \int_{-\infty}^{\infty} \prod_{i=1}^{I} P_{ik}(\theta)\phi(\theta)d\theta, \qquad (8)$$

where $\phi(\cdot)$ is the standard normal density. Jointly considering all I = 6 rating criteria as items, we obtain an estimate of the parameter vector $\widehat{\boldsymbol{B}}$, and an estimate $\hat{\theta}$ of the proficiency of a given ratee, giving a linearized transformation of the original scores.

The use of polychoric correlation during our factor analysis procedure accounts for attenuation of latent variable correlations, and it can be shown that the parameter vector $\boldsymbol{B}$ can be estimated using two-stage factor analysis using polychoric correlation (Samejima, 1969). However, the least-squares solution $\widehat{\boldsymbol{F}} = \left(\hat{\boldsymbol{L}}^{\boldsymbol{T}}\,\hat{\boldsymbol{L}}\right)^{-1}\hat{\boldsymbol{L}}^{\boldsymbol{T}}\,(\boldsymbol{PRR} - \hat{\boldsymbol{\mu}})$ fails to capture potential nonlinearities in the relationship between the latent space and the item scores, which is captured in the estimation of $\hat{\theta}$.

Appendix C. Supplemental Tables

Table C1. Inter-rater reliability with and without relationship controls

	Relationship Controls?	Percentage of Total Variability		Total Variability	Inter-Rater Reliability		
		Applicant	Residual		IRR Est.	LCI	UCI
Factor	Yes	30%	70%	0.87	0.30	0.27	0.32
	No	28%	72%	0.89	0.28	0.26	0.30
Theta	Yes	29%	71%	0.86	0.29	0.27	0.32
	No	28%	72%	0.89	0.28	0.26	0.30
Overall	Yes	31%	69%	1.36	0.31	0.28	0.33
	No	29%	71%	1.41	0.29	0.27	0.31
Engagement	Yes	28%	72%	1.37	0.28	0.25	0.30
	No	27%	73%	1.40	0.27	0.25	0.29
Instruction	Yes	29%	71%	1.36	0.29	0.26	0.31
	No	27%	73%	1.40	0.27	0.25	0.29
Management	Yes	31%	69%	1.54	0.31	0.29	0.33
	No	31%	69%	1.58	0.31	0.29	0.33
Diverse	Yes	23%	77%	1.28	0.23	0.21	0.25
	No	22%	78%	1.33	0.22	0.2	0.25
Interpersonal	Yes	28%	72%	1.37	0.28	0.26	0.30
	No	27%	73%	1.40	0.27	0.24	0.29
Challenges	Yes	27%	73%	1.38	0.27	0.24	0.29
	No	26%	74%	1.42	0.26	0.24	0.28

Notes: Each outcome represents a separate regression model estimated using equation (3), controlling for rater internal status and relationship effects.

Table C2. Inter-rater reliability by applicant type: novice versus experienced

		Applicant Type		Percentage of Total Variability		Total Variability	Inter-Rater Reliability			Novice - Experienced		
		b	SE(b)	Applicant	Residual		IRR Est.	LCI	UCI	Dif. Est.	LCI	UCI
Factor	Novice	-0.10	0.03	26%	74%	0.83	0.26	0.22	0.30	0.05	0.00	0.09
	Experienced	(Ref)		30%	70%	0.88	0.30	0.28	0.33			
Theta	Novice	-0.11	0.03	23%	77%	0.80	0.23	0.19	0.28	0.07	0.02	0.12
	Experienced	(Ref)		30%	70%	0.88	0.30	0.28	0.33			
Overall	Novice	-0.08	0.04	28%	72%	1.31	0.28	0.23	0.33	0.03	-0.02	0.08
	Experienced	(Ref)		31%	69%	1.37	0.31	0.28	0.33			
Engagement	Novice	-0.13	0.04	25%	75%	1.32	0.25	0.21	0.29	0.03	-0.02	0.08
	Experienced	(Ref)		28%	72%	1.37	0.28	0.26	0.31			
Instruction	Novice	-0.16	0.04	24%	76%	1.28	0.24	0.19	0.28	0.05	0.01	0.10
	Experienced	(Ref)		29%	71%	1.37	0.29	0.27	0.32			
Management	Novice	-0.13	0.04	25%	75%	1.42	0.25	0.21	0.30	0.70	0.01	0.12
	Experienced	(Ref)		32%	68%	1.42	0.32	0.30	0.34			
Diverse	Novice	-0.11	0.04	22%	78%	1.26	0.22	0.17	0.26	0.02	-0.03	0.07
	Experienced	(Ref)		23%	77%	1.29	0.23	0.21	0.26			
Interpersonal	Novice	-0.04	0.04	23%	77%	1.29	0.23	0.19	0.28	0.05	0.00	0.10
	Experienced	(Ref)		28%	72%	1.38	0.28	0.27	0.31			
Challenges	Novice	-0.13	0.04	22%	78%	1.30	0.22	0.18	0.27	0.05	0.00	0.10
	Experienced	(Ref)		28%	72%	1.39	0.28	0.25	0.30			

Notes: Each outcome represents a separate regression model estimated using equation (6), controlling for rater internal status and relationship effects (coefficient estimates not shown). Differences between Inter-Rater Reliability by applicant type are calculated within bootstrap iteration to ensure comparability.

Table C3. Inter-rater reliability by applicant type: internal versus external

		Applicant Type		Percentage of Total Variability		Total Variability	Inter-Rater Reliability			Novice - Experienced		
		b	SE(b)	Applicant	Residual		IRR Est.	LCI	UCI	Dif. Est.	LCI	UCI
Challenges	Internal	-0.07	0.05	33%	67%	0.50	0.33	0.28	0.38	0.03	0.08	0.13
	External	(Ref)		25%	75%	0.33	0.25	0.22	0.28			
Management	Internal	-0.09	0.05	37%	63%	0.62	0.37	0.32	0.41	0.08	0.03	0.13
	External	(Ref)		29%	71%	0.44	0.29	0.26	0.33			
Diverse	Internal	-0.05	0.04	29%	71%	0.42	0.29	0.25	0.34	0.10	0.05	0.15
	External	(Ref)		19%	81%	0.23	0.19	0.17	0.22			
Interpersonal	Internal	-0.09	0.05	31%	69%	0.45	0.31	0.26	0.35	0.05	0.00	0.10
	External	(Ref)		26%	74%	0.34	0.26	0.23	0.28			
Engagement	Internal	-0.09	0.05	35%	65%	0.53	0.35	0.31	0.40	0.10	0.05	0.15
	External	(Ref)		25%	75%	0.33	0.25	0.23	0.28			
Instruction	Internal	-0.04	0.05	37%	63%	0.57	0.37	0.33	0.42	0.12	0.07	0.17
	External	(Ref)		25%	75%	0.33	0.25	0.23	0.28			
Overall	Internal	-0.08	0.05	36%	64%	0.54	0.36	0.32	0.40	0.03	0.08	0.13
	External	(Ref)		28%	72%	0.36	0.28	0.25	0.31			
Factor	Internal	-0.05	0.04	37%	63%	0.36	0.37	0.33	0.41	0.10	0.05	0.15
	External	(Ref)		27%	73%	0.22	0.27	0.23	0.30			
Theta	Internal	-0.05	0.04	36%	64%	0.35	0.36	0.32	0.40	0.10	0.04	0.14
	External	(Ref)		27%	73%	0.23	0.27	0.23	0.30			

Notes: Each outcome represents a separate regression model estimated using equation (6), controlling for rater internal status and relationship effects (coefficient estimates not shown). Differences between Inter-Rater Reliability by applicant type are calculated within bootstrap iteration to ensure comparability.