NATIONAL
CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

CALDER

AIR

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

Dan Goldhaber
Cyrus Grout
Malcom Wolff
Patricia Martinkova

WORKING PAPER No.237-0620 • June 2020

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

Dan Goldhaber
American Institutes for Research/CALDER
University of Washington

Cyrus Grout
University of Washington

Malcolm Wolff
University of Washington

Patricia Martinkova
Charles University
Institute of Computer Science of the Czech Academy of Sciences

Contents

Contents i

Acknowledgments ii

Abstract iii

1. Introduction 1

2. Background Literature: Hiring Preferences and Screening Applicants 3

3. The Application Process and the Collection of Reference Ratings 5

4. Data 8

5. Empirical Approach 10

5.1 Exploratory Factor Analysis of Distinct Traits Captured by Reference Ratings Survey 10

5.2 Inter-Rater Reliability 12

6. Results 15

6.1 *Factor Analysis of Reference Ratings* 15

6.2 *Inter-Rater Reliability* 17

7. Discussion and Conclusions 19

References 23

Tables 28

Figures 33

Appendix 39

Polychoric Correlation 39

Graded Response Modeling 40

Supplementary Tables 41

Acknowledgments

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. 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.

CALDER working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street NW, Washington, DC 20007
202-403-5796 • www.caldercenter.org

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

Dan Goldhaber, Cyrus Grout, Malcolm Wolff, Patricia Martinkova
CALDER Working Paper No. 237-0620
June 2020

Abstract

There is growing interest in using measures of teacher applicant quality to improve hiring decisions, but the statistical properties of such measures are poorly understood. We present evidence on structured ratings solicited from teacher applicants' references. We find that the reference ratings capture only one underlying dimension of applicant quality, which may indicate a need to broaden the range of questions posed to professional references. 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 comparable applicant assessment tools.

Keywords: teacher applicants, teacher hiring, inter-rater reliability

1. Introduction

When hiring teachers, school principals (or other district hiring officials) are certainly selecting teacher applicants on what they think are multiple dimensions of quality. Principals, for instance, report seeking to hire experienced teachers with good classroom management skills, 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). These types of skills or attributes may be thought of as different dimensions of teacher quality. And while some of them could be associated with readily objective and quantifiable attributes, such as teacher experience and performance on licensure tests, others may not be. Hiring officials' assessments of traits such as communication or classroom management skills, cultural competence, or caring are more likely to be based on subjective assessments of teacher applicant materials and to be more difficult to quantify. This raises the question, to what extent can school systems collect *meaningful* and *reliable* 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 an expanding menu of commercially available instruments that school districts can use to rate or pre-screen teacher applicants.[1] Yet there is little evidence on the dimensionality and reliability of these measures of applicant quality – factors that affect their

[1] 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).

ability to inform hiring decisions. [2] This is surprising and represents an important gap in the literature.

Our research focuses on the potential to solicit better information about job applicants from their professional references. The practice of collecting letters of recommendation from job applicants' references is widespread (Aamodt et al., 1993; Salgado, 2001) and as discussed in **Section 2**, there is some evidence that information provided by references is predictive of subsequent teacher performance (Goldhaber et al., 2017). Incorporating the collection of structured reference ratings into the teacher application process is a potentially low-cost, easy-to-implement means of enhancing the applicant information available to hiring officials. This stands in contrast to the centralized screening systems studied by Jacob et al. (2018) and Bruno and Strunk (2019), which require one-on-one interactions with district administrators, and can be quite costly.

We present evidence from a survey completed by the references of applicants (those who write letters of recommendation for the applicant) to a medium-sized urban school district (henceforth, the District). The survey, the development of which is described in **Section 3**, is designed to solicit information about various dimensions of applicant quality; specifically, references were 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", and that the prevalence of cheerleading varies according to rater type (e.g., for principals

[2] Surveys like the one we employ are widely used to collect information about applicants to graduate schools and medical residency programs (e.g., Girzadas et al., 1998; Liu et al., 2009; Oliveri et al., 2017; McCaffrey et al., 2018), and there is some empirical work on the inter-rater reliability of ratings in these contexts (see, in particular, McCaffrey et al., 2018),

compared to colleagues). Regarding dimensionality, we find that the reference survey captures only one underlying dimension of applicant quality. This may indicate a need to broaden the range of questions posed to PRs and/or that the number questions posed could be reduced without losing information. Point estimates of inter-rater reliability range between 0.23 and 0.31 depending on the evaluation criterion and are significantly higher for experienced applicants relative to novice applicants and for applicants with prior experience in the District relative to applicants with out-of-district teaching experience only. These findings may provide guidance for how to improve the rating process in the future.

2. Background Literature: Hiring Preferences and Screening Applicants

Most of the research on the teacher traits valued by hiring officials assesses the extent to which different characteristics of teacher applicants determine teacher selection. Ballou (1996), for instance, used national data about individuals who applied for a teaching position to examine the extent to which hiring officials value individuals who graduate from more selective colleges or have math or science academic majors. He found little difference in the probability of employment as a teacher according to those measures of academic quality, and thus concluded that hiring officials were not overly concerned with the academic qualifications of prospective teachers. Boyd et al. (2011) studied teachers' applications to transfer from one school to another; this avoids the problem of not being able to distinguish teachers' decisions to apply from principals' decisions to hire. In contrast to Ballou (1996), they found that applicants' academic qualifications were predictive of hiring decisions, a finding bolstered by a more recent analysis looking at the entry of new elementary teachers into the workforce (Boyd et al., 2013).

The studies above utilize data that is collected in surveys or available in administrative records, i.e. it is *readily quantifiable* in the sense that it is based on objective measures. In contrast, using information collected from principal interviews, Harris et al. (2010) found that hiring officials value some of these readily quantifiable attributes but also rely on more subjective judgements, such as an applicant's anticipated organizational fit. Several more recent studies analyzed a broader range of teacher applicant information than is typically available in administrative data. Specifically, these studies looked at information about applicants collected during the hiring process with a focus on the predictive validity of measures of applicant quality. These measures include: subjective assessments that are part of screening rubrics used to determine which applicants advance in the hiring process (Goldhaber et al., 2017; Bruno and Strunk, 2018; Jacob et al., 2018); ratings on commercially available applicant selection tools (Jacob et al., 2018); and measures of work experience relevance and tenure history derived from resumes (Sajjadiani et al., 2018). All these studies found that measures of applicant quality generated during the hiring process are, to various degrees, predictive of subsequent teacher outcomes.

Interestingly, despite the ubiquity of using professional references as a means of screening job applicants (Salgado, 2001), there is relatively little evidence about the degree to which those references distinguish between different dimensions of applicant quality or whether different PRs tend to agree about applicant quality. McCarthy and Goffin (2001) looked at 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. Outside of the job application context, some new research has looked at the properties of *personal* reference ratings in the context of applicants to graduate school programs (e.g., Oliveri et al., 2017; McCaffrey et al., 2018).

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.[3] Applicants were scored 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 test differences in inter-rater reliability and found that the within-school inter-rater reliability of the summative rating was higher for applicants from within the district (0.51) than for those from outside the district (0.42). They also found that the reliability of ratings on some dimensions of applicant quality were quite low – on "cultural competency" for instance it was only 0.35 for internal applicants and 0.33 for external applicants.

3. The Application Process and the Collection of Reference Ratings

The first step for individuals wishing to apply for a job in the study District is to create an applicant profile in the online applicant management system. In the profile, the 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 District directing them to an online submission form.

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

That form records the letter writer's name, e-mail address, and relationship to the applicant.[4] Having completed a profile, applicants can apply to any number of specific job postings.[5]

To narrow the pool of applicants who will be more closely considered for a position, school principals request that HR provide a truncated list of applicants (typically the top 7 to 10) 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. Applicants receiving top ratings are invited for in-person interviews.

In June 2015, as part of a collaboration with the District designed to study and improve teacher hiring practices, we began collecting structured assessments of applicants from their references. Following the submission of a letter of recommendation, references are redirected to an online survey where they are asked to rate applicants relative to their 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:

[4] 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.

[5] Prior to openly listing a job posting, the District typically lists the position internally and in accordance with protocols outlined in the district's collective bargaining agreement, considers the two most senior applicants (in terms of district tenure) to position. For the purposes of this study, we only observe job postings that were openly listed such that any applicant could apply.

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)

[FIGURE 1 ABOUT HERE]

[TABLE 1 ABOUT HERE]

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 the District. These competencies, described in **Table 1**, are: classroom management, instructional skills, interpersonal skills, challenging students, student engagement, and working with diverse groups of students.[6]

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

[6] Three of these competencies were demonstrated in previous work to be significantly predictive of teacher outcomes (Goldhaber et al., 2017): classroom management, instructional skills, and interpersonal skills. Two others, student engagement and challenging students, are selected on the basis of evidence on Ron Ferguson's Tripod survey instrument 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 the District.

pattern henceforth referred to as "cheerleading".[7] 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. A plurality of applicants (41%) have three reference ratings, 18% have four, 9% five, and 4% have six.[8] 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

[7] Leising, Erbs, and Fritz (2010), for instance, find that in studies using ratings of personality, raters who liked their ratees better have been found to rate them more positively (e.g., as being more extroverted, agreeable, open, conscientious, and less neurotic).

[8] 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.

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.[9] 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 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 (i.e., the 3,601 observations referenced above).

As illustrated in **Figure 2**, which shows the distribution of reference ratings for the "Overall" criterion, the distribution of ratings across rater type varies considerably.[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

[9] A data crosswalk generated by the District's hiring database was used to link reference IDs to the reference ratings data. These 8 reference ratings were missing from that crosswalk and could not be tracked down manually.
While confidential letters of recommendation submitted through the District's hiring pipeline have a letterhead with a field for relationship type, some references chose to submit letters on their own letterhead.

[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.

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.

[FIGURE 2 ABOUT HERE]

Descriptive statistics for the analytic sample are presented in **Table 2**. Applicants tend to be experienced; only 11% report no professional teaching experience, while 16% have teaching experience in the District. 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".

[TABLE 2 ABOUT HERE]

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.

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 raw 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. Using these correlations, we identify the latent characteristics underlying references’ judgements of applicant quality. Formally, the k centered reference ratings criteria can be described by,

$$PRR_k - \mu_k = l_{k1}F_1 + \cdots + l_{kD}F_D + \varepsilon_k, \qquad (1)$$

for D latent factors F_d and mean zero error terms ε_k. Equation (1) 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 factors and weights derived from the factor analysis will be used to generate a summative ratings measure: *PR Factor*. A limitation of this approach is that it assumes differences of equal magnitude between each rating level, which is not highly-credible here. For instance, the difference between a rating of “Among the best (top 1%)” and “Excellent (top 10%)” may be small relative to the difference between “Excellent (top 10%)” and “Average.” As a robustness check, we also generate a second summative ratings measure ($Theta$) derived from the graded response model (GRM) described in the **Appendix**. The GRM, introduced by Samejima (1969), addresses a limitation we face in analyzing the ratings data – that assuming differences of equal magnitude between each rating level may not be valid. As a summative measure of applicant quality, the

GRM model allows us to relax the assumption that the distances between ratings levels are constant.

5.2 Inter-Rater Reliability

In the context of this paper, inter-rater reliability measures the extent to which different references agree about the quality 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. Each characteristic of the measurement situation that a decision maker would be indifferent to (e.g., the occasion, the rater, or the item/criterion) is a potential source of error and is called a facet of a measurement. In order to evaluate the generalizability and dependability of the ratings, as many facets of measurement error are isolated and estimated as is feasible.

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 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.

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, District 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 District 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 in the following mixed effect model:[11]

$$PRR_{ij} = \mu + \alpha_1{}'\boldsymbol{r}_{ij} + \alpha_2\, s_j + A_i + \varepsilon_{ij}\,, \qquad (2)$$

where PRR_{ij} is the 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 employee of the District, 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. A significant coefficient in the vector $\hat{\alpha}_1$ or on $\hat{\alpha}_2$ suggests variation driven by reference type. We then estimate the contribution of variance from applicants in each group using the variance decomposition model:

$$\sigma_{PRR}^2 = \sigma_A^2 + \sigma_\varepsilon^2, \qquad (3)$$

where σ_A^2 represents the systematic error-free variance among scores 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 the Appendix):

$$IRR = \frac{\sigma_A^2}{\sigma_A^2+\sigma_\varepsilon^2}\ . \qquad (4)$$

Equation (4) 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 District ("internal applicants") to applicants with teaching experience outside of the District ("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 the District 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 (2)** to differ by group by assuming the following mixed-effect model:

$$PRR_{ijg} = \mu + \alpha_1{}'\boldsymbol{r}_{ij} + \alpha_2\, s_j + \alpha_3 p_i + A_{ig} + \varepsilon_{ijg}\,, \qquad (5)$$

Where PRR_{ij} 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 the District, 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 (3)** then becomes:

$$\sigma^2_{PRR,g} = \sigma^2_{Ag} + \sigma^2_{\varepsilon}, \qquad (6)$$

And the inter-rater reliability for these groups is then calculated using **equation (7)**:

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

We use bootstrap procedures to calculate confidence intervals around the point estimates for inter-rater reliability and also 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 3**; we suppress output for factors beyond Factor 1 because they explain so little covariation. We find that each reference ratings measure loads onto Factor 1 and that the loadings are of similar magnitude (between 0.89 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 (5.11) and the small eigenvalues of subsequent Factors (see **Figure 3**). In fact, Factor 1 explains 96% of cumulative variation.[12]

[TABLE 3 ABOUT HERE]

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.[13]

[FIGURE 3 ABOUT HERE]

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.

[12] As an additional measure of 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 .81 across all regressions and ranges from .74 (regressing "working with diverse groups of students" on "classroom management") to .90 (regressing "instructional skills" on "classroom management"). In regressions with multiple covariates, we find all coefficients are relatively similar, with diverse displaying the most substantial deviations.

[13] 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**.

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 4**, 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.

[TABLE 4 ABOUT HERE]

As noted above, we use the factor loadings presented in **Table 3** to generate a "PR Factor" rating. 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*) to address potential problems arising from imposing a uniform distance between rating levels on the reference ratings data. The GRM measure 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 (2)** for each rating criterion and our two composite measures, *PR Factor* and *Theta*, are presented in **Table 5**. 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 PR 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 – between 43% and 60% of a standard deviation lower than colleagues.. 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 (2) (5)**). Estimates of inter-rater reliability ***unadjusted*** by rater type are presented in **Table A2** in the appendix.

[TABLE 5 ABOUT HERE]

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, in general fall within the margin of error of one another. The exception is for the criterion "Working with Diverse Groups of Students," which has a far lower point estimate of 0.23. These findings are similar to those reported in Martinková et al. (2018) who find that the inter-rater reliability of ratings of teacher applicants on the criterion "Cultural Competency" is lower than for other criteria. This suggests that the lower ratings on "working with diverse groups of students" criterion are likely due to a general difficulty that educators have in agreeing to what it means to be effective at working with diverse groups of students.

[FIGURE 4 ABOUT HERE]

In **Figure 5** and **Figure 6** (also see Supplementary Tables **A2** and **A3** in the appendix**)**, 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 the District to that of external applicants who report teaching experience outside of the District 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]

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]

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 applicants 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 indeed multiple underlying traits 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. For example, 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 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 appropriate 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 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. And Rust and Golombok (2009) notes 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.[14]

A potential limitation to the inter-rater reliability of the reference ratings studied here is that the raters are not in a position to receive training on how to rate applicants. 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. A second limitation is that raters are likely to have known a particular applicant under different circumstances or during different periods of time (e.g., as a university student versus as a fellow teacher), meaning that in some cases they are forming judgements about the applicant using different sets of information.

A lower level of inter-rater reliability may be acceptable in the context of professional reference ratings (vs. 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. That

[14] 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).

said, 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), but also that we have very limited evidence on the reliability of other means of gathering subjective information about applicants (e.g., assessments of sample teaching lessons).

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 policy actions 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. Fewer of these questions could be asked without losing information, or different questions could be developed to try to capture other dimensions of applicant quality.

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

Ballou, D. (1996). Do Public Schools Hire the Best Applicants? *The Quarterly Journal of Economics*, *111*(1), 97–133.

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

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., and Wyckoff, J. (2011). The role of teacher quality in retention and hiring: Using applications to transfer to uncover preferences of teachers and schools. *Journal of Policy Analysis and Management*, *30*(1), 88–110. doi:10.1002/pam

Boyd, D., Lankford, H., Loeb, S., and Wyckoff, J. (2013). Analyzing the Determinants of the Matching of Public School Teachers to Jobs: Disentangling the Preferences of Teachers and Employers. *Journal of Labor Economics*, *31*(1), 83–117. doi:10.1086/666725

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 &amp; Evaluation, 18(8). *Practical Assessment, Research & Evaluation*, *18*(8), 1–14.

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

Girzadas, D. V, Harwood, R. C., Dearie, J., and Garrett, S. (1998). A comparison of standardized and narrative letters of recommendation. *Academic Emergency Medicine*, *5*(11), 1101–1104. doi:10.1111/j.1553-2712.1998.tb02670.x

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

Leising, D., Erbs, J., and Fritz, U. (2010). The letter of recommendation effect in informant ratings of personality. *Journal of Personality and Social Psychology*, *98*(4), 668–682.

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., 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

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

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

Tables

Table 1. Description of Criteria for References' Ratings of Applicants

Criterion	Description
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
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
Challenges Students	• Sets high expectations and holds students accountable

Table 2. Descriptive Statistics

	All Raters	Colleague	Instr. Coach/ Dept. Chair	Cooperating Teacher	Principal/ Other Sup.	University Supervisor	Other
Ratings							
Engagement	4.65	4.92	4.68	4.53	4.42	4.64	4.83
	(1.18)	(1.07)	(1.18)	(1.23)	(1.25)	(1.12)	(1.09)
Instruction	4.53	4.82	4.57	4.39	4.31	4.52	4.66
	(1.18)	(1.09)	(1.19)	(1.22)	(1.24)	(1.10)	(1.09)
Management	4.44	4.73	4.50	4.16	4.29	4.35	4.58
	(1.26)	(1.17)	(1.23)	(1.29)	(1.31)	(1.18)	(1.17)
Diverse	4.73	4.99	4.77	4.58	4.52	4.69	4.91
	(1.15)	(1.04)	(1.16)	(1.19)	(1.19)	(1.09)	(1.09)
Interpersonal	4.73	4.97	4.76	4.71	4.46	4.85	4.87
	(1.18)	(1.09)	(1.21)	(1.21)	(1.26)	(1.04)	(1.11)
Challenges	4.52	4.81	4.52	4.35	4.32	4.47	4.67
	(1.19)	(1.10)	(1.2)	(1.21)	(1.24)	(1.13)	(1.12)
Overall	4.52	4.83	4.56	4.45	4.25	4.48	4.63
	(1.19)	(1.08)	(1.17)	(1.22)	(1.25)	(1.09)	(1.11)
Applicants							
Teaching Experience	6.41	8.34	7.47	1.67	8.20	1.87	5.07
	(7.19)	(7.21)	(7.31)	(2.95)	(7.65)	(3.81)	(6.7)
Female	0.69	0.70	0.80	0.68	0.70	0.67	0.66
	(0.46)	(0.46)	(0.41)	(0.48)	(0.46)	(0.48)	(0.48)
Internal	0.16	0.17	0.21	0.13	0.19	0.12	0.13
	(0.37)	(0.38)	(0.41)	(0.34)	(0.39)	(0.33)	(0.33)
Novice	0.11	0.04	0.06	0.28	0.06	0.30	0.12
	(0.32)	(0.20)	(0.23)	(0.45)	(0.24)	(0.46)	(0.33)
Raters							
Internal Rater	0.15	0.13	0.15	0.26	0.18	0.01	0.10
	(0.35)	(0.33)	(0.36)	(0.44)	(0.39)	(0.07)	(0.30)
Observations	10,763	2,792	454	1,238	3,598	979	1,702

Notes: Descriptive statistics of reference ratings, applicant characteristics, and rater internal status by reference-applicant relationship. 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.

Table 3. Initial Factor Extraction

	Factor 1
Engagement	0.96
Instruction	0.95
Management	0.92
Diverse	0.90
Interpersonal	0.86
Challenges	0.94
Cumulative Variation Explained	0.96
Eigenvalue	5.11

Notes: Factor weights associated with each rating criteria for the primary factor, which explains 96% of the variation across items.

Table 4: Correlations of Ratings Criteria and Overall Measures

	Factor	Theta	Overall	Engagement	Instruction	Management	Diverse	Interpersonal	Challenges
Factor	1								
Theta	0.98	1							
Overall	0.92	0.91	1						
Engagement	0.94	0.93	0.92	1					
Instruction	0.93	0.93	0.92	0.90	1				
Management	0.91	0.90	0.89	0.88	0.88	1			
Diverse	0.89	0.86	0.85	0.85	0.83	0.81	1		
Interpersonal	0.87	0.83	0.86	0.83	0.81	0.78	0.80	1	
Challenges	0.93	0.92	0.90	0.90	0.91	0.87	0.84	0.79	1

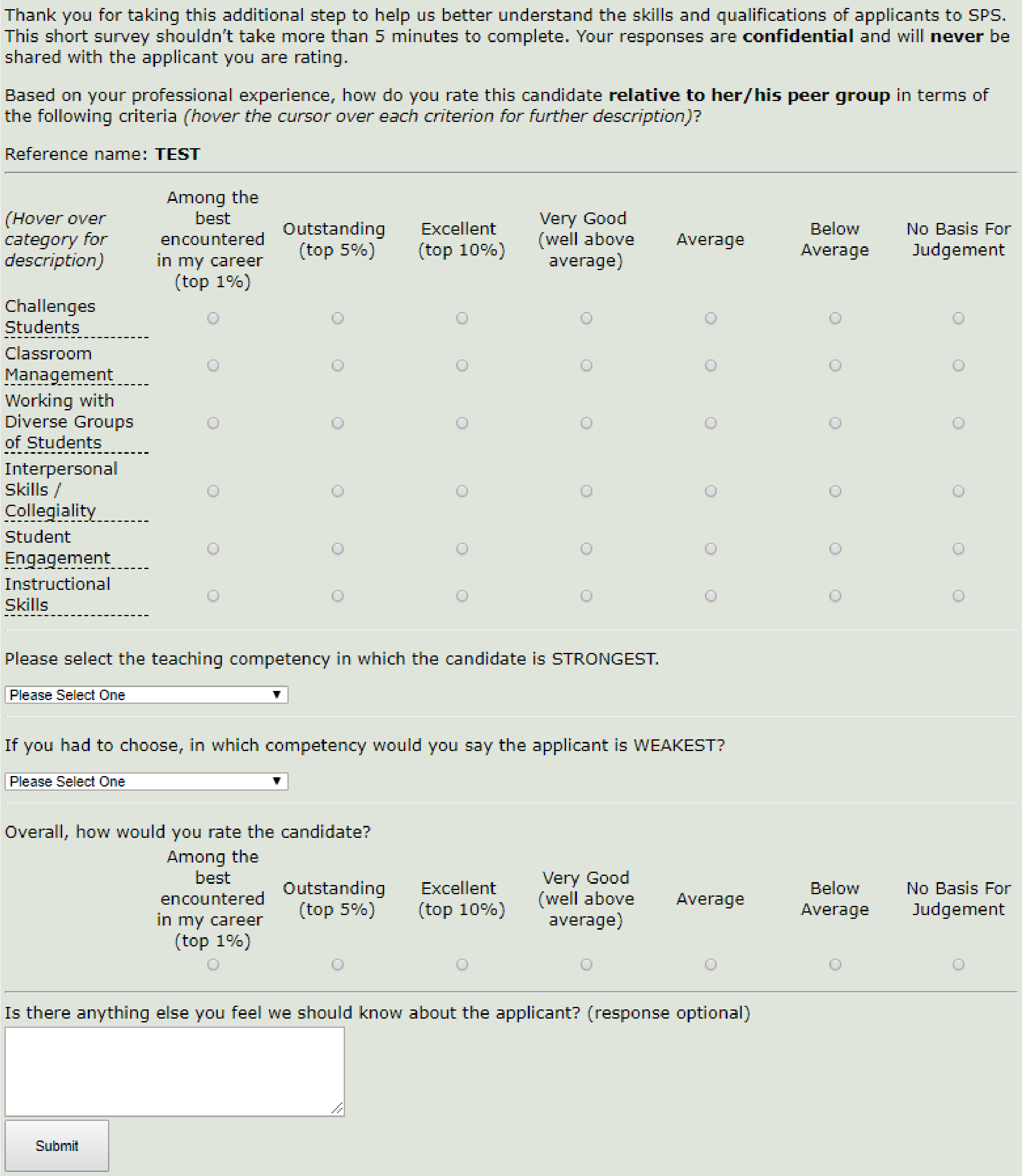
Notes: Coefficients displayed are calculated using polychoric correlations on all non-missing criteria from 10,763 ratings and 3,070 applicant clusters.

Table 5. Mixed effect models with rater-type fixed effects

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Factor	Theta	Overall	Engagement	Instruction	Management	Diverse	Interpersonal	Challenges
Relationship Type	Colleague	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)
	Instructional Coach/Dept. Chair	-0.242***	-0.260***	-0.323***	-0.279***	-0.295***	-0.251***	-0.263***	-0.260***	-0.337***
		(0.044)	(0.044)	(0.055)	(0.056)	(0.056)	(0.06)	(0.056)	(0.056)	(0.057)
	Cooperating Tchr.	-0.274***	-0.283***	-0.294***	-0.293***	-0.305***	-0.448***	-0.341***	-0.222***	-0.361***
		(0.031)	(0.031)	(0.039)	(0.04)	(0.04)	(0.043)	(0.039)	(0.04)	(0.04)
	Principal	-0.436***	-0.435***	-0.605***	-0.504***	-0.531***	-0.457***	-0.488***	-0.536***	-0.506***
		(0.022)	(0.022)	(0.028)	(0.029)	(0.029)	(0.031)	(0.028)	(0.028)	(0.029)
	Univ. Supervisor	-0.207***	-0.233***	-0.325***	-0.239***	-0.239***	-0.313***	-0.262***	-0.105***	-0.293***
		(0.034)	(0.034)	(0.042)	(0.043)	(0.043)	(0.047)	(0.043)	(0.043)	(0.044)
	Other	-0.099***	-0.100***	-0.155***	-0.062***	-0.098***	-0.100***	-0.059***	-0.090***	-0.104***
		(0.028)	(0.028)	(0.035)	(0.037)	(0.037)	(0.042)	(0.036)	(0.035)	(0.038)
	Internal Rater	-0.044	-0.042	-0.083***	-0.080***	-0.075***	-0.052	-0.005	0.002	-0.064**
		(0.028)	(0.028)	(0.035)	(0.036)	(0.036)	(0.039)	(0.034)	(0.035)	(0.036)
	Intercept	4.226***	0.215***	4.840***	4.923***	4.810***	4.728***	4.990***	4.981***	4.813***
		(0.019)	(0.019)	(0.024)	(0.025)	(0.025)	(0.027)	(0.024)	(0.024)	(0.025)
	Applicant variance	0.2327	0.2353	0.383	0.3463	0.3553	0.4364	0.2683	0.3436	0.3374
	Residual variance	0.6308	0.6239	0.9656	1.0096	0.9956	1.0937	1.0089	1.0138	1.0358
	Applicant clusters	3,070	3,070	3,070	3,070	3,070	3,070	3,070	3,070	3,070
	Observations	10,763	10,763	10,763	10,763	10,763	10,763	10,763	10,763	10,763

Figures

Figure 1. Professional Reference Survey Form



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?

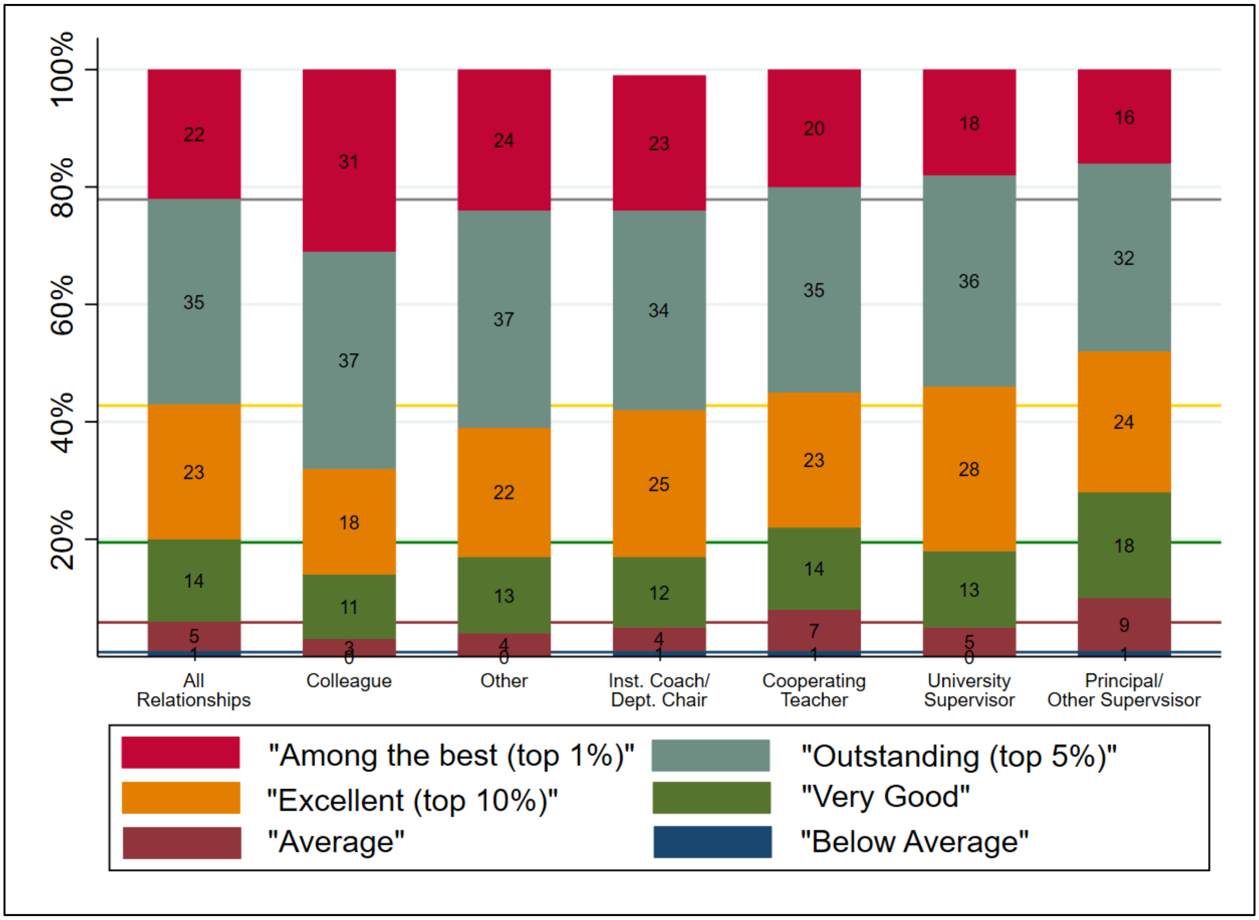
Please Select One ▼

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
○	○	○	○	○	○	○

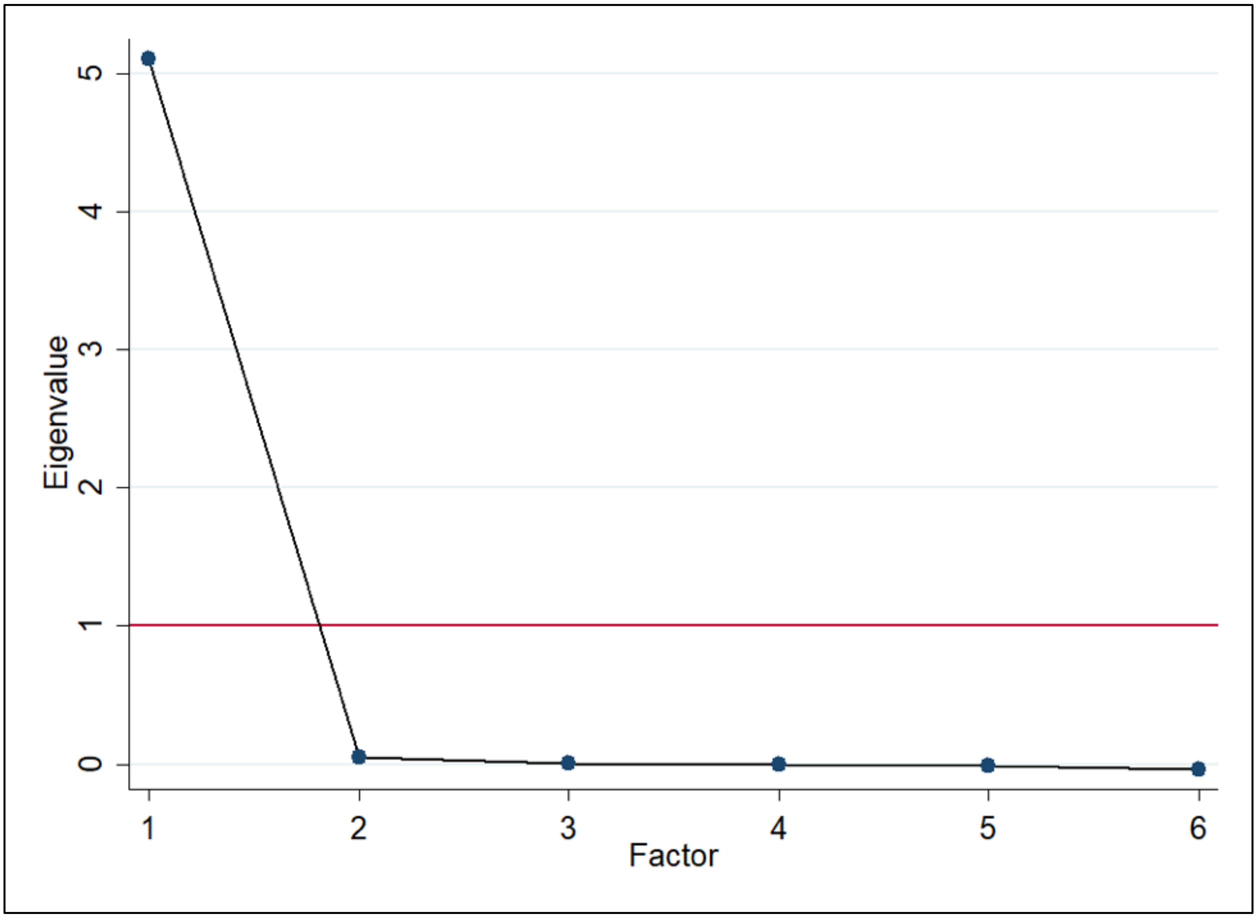
Is there anything else you feel we should know about the applicant? (response optional)

Figure 2: Distribution of Ratings on "Overall" Criterion by Rater Type



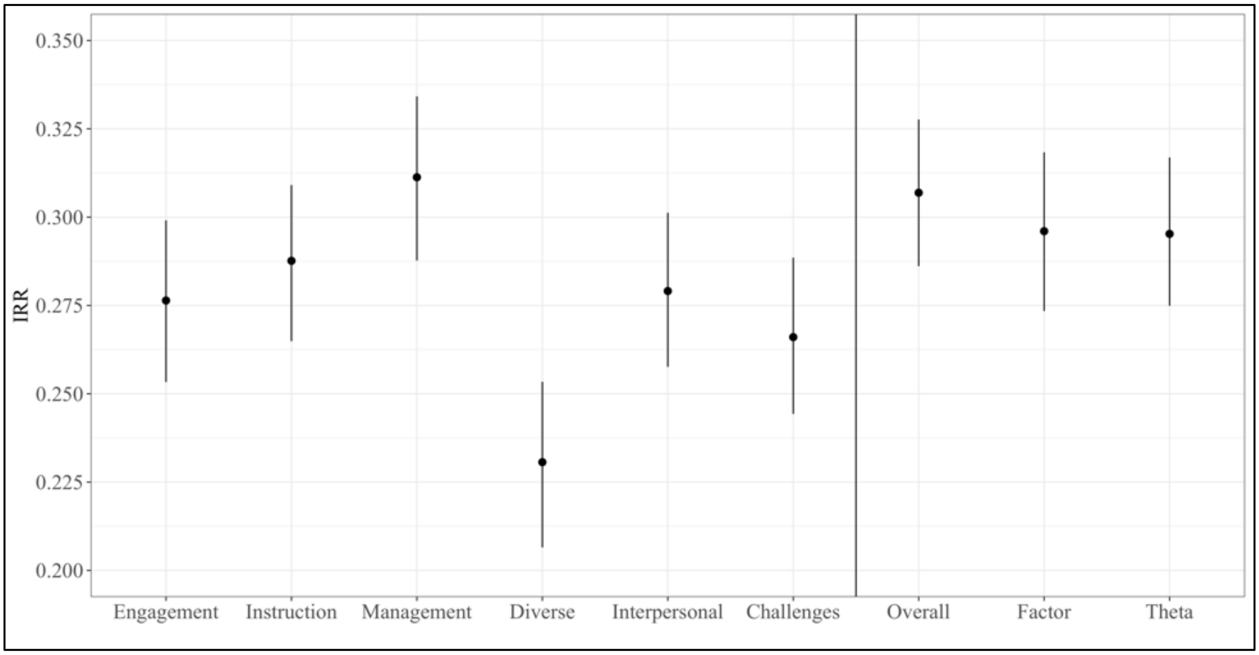
Notes: Distribution of ratings by applicant-rater relationship type (N = 10,763). Note that ratings criteria for which the reference indicated a rating of "no basis for judgement" are treated as missing values. This results in 356 missing values for the student engagement criterion, 457 for instructional 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 3. Scree plot of eigenvalues



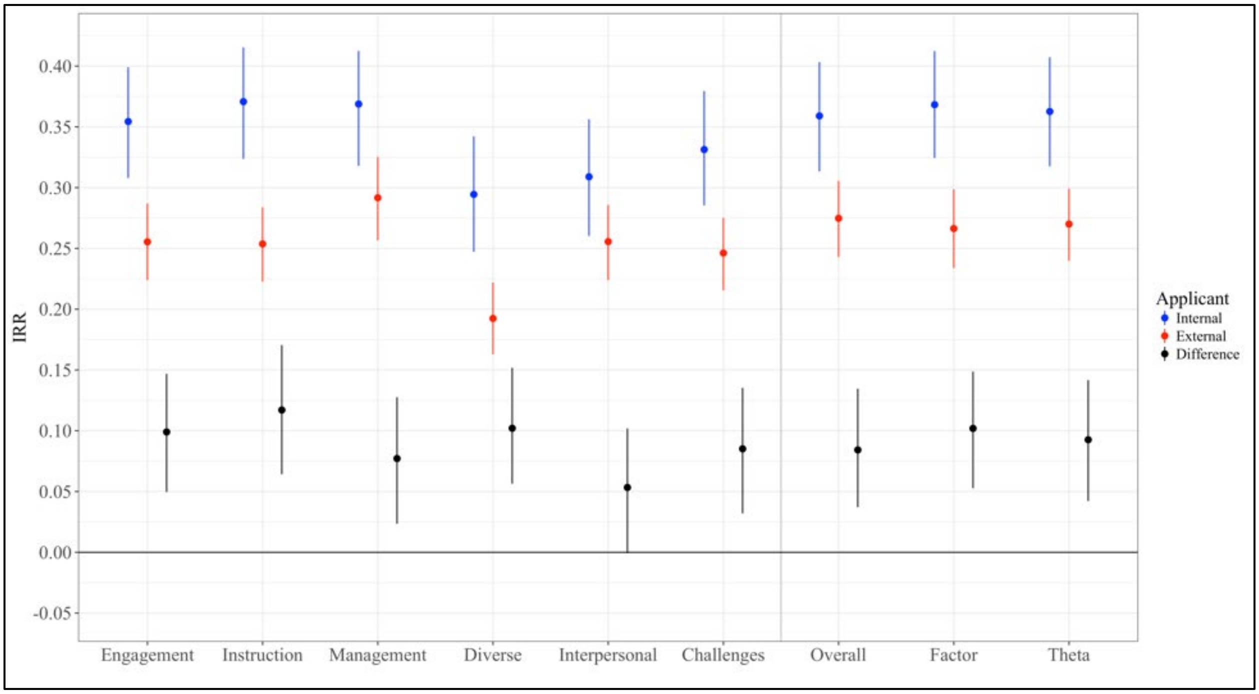
Notes: Scree plot of eigenvalues for six factors generated by the rating matrix including categories Engagement, Instruction, Management, Diverse, Interpersonal, and Challenges. As a rule of thumb, eigenvalues above 1 are taken to be unique factors.

Figure 4: Inter-Rater Reliability by Rating Category



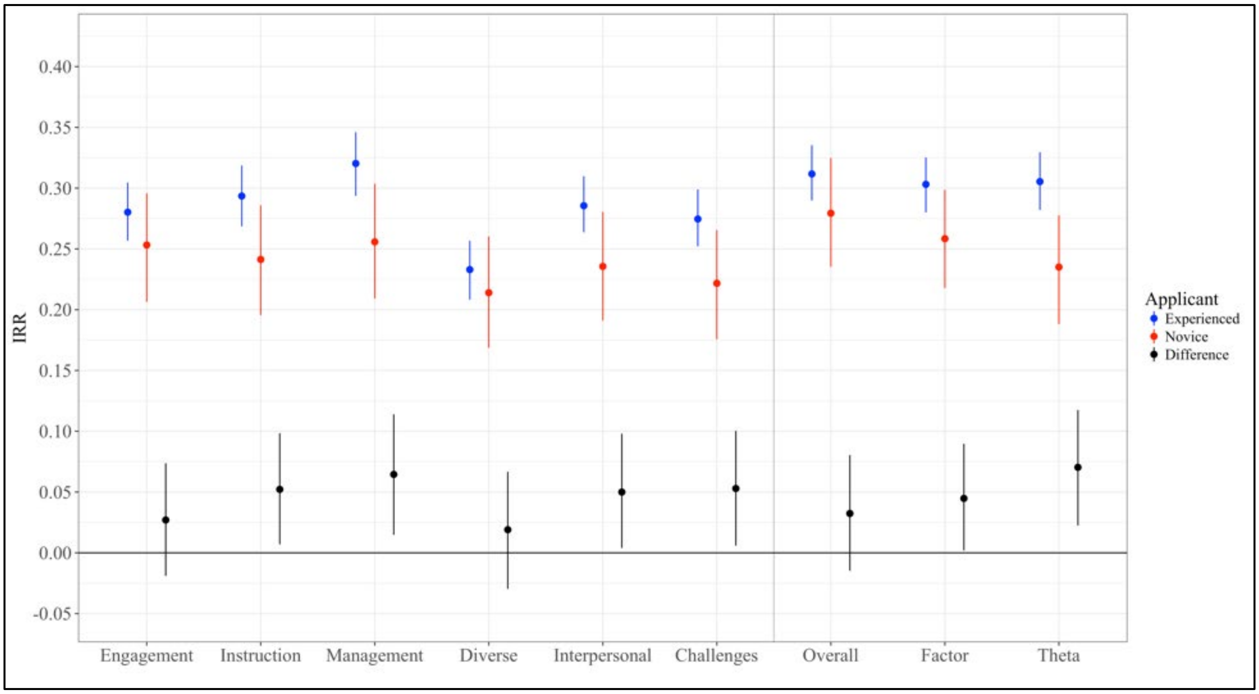
Notes: Point estimates of IRR across rating category, “overall” rating, the generated PR Factor, and *Theta*, using 3,601 applicant-years across 10,763 ratings. Confidence intervals are generated using parametric bootstrap with 1,000 replications.

Figure 5: Inter-Rater Reliability by Rating Category and Applicant Type (Internal/External)



Notes: Point estimates of IRR by applicant internal/external status across rating category, including "overall", the generated PR Factor, and *Theta*, using 3,601 applicant-years across 10,763 ratings (15% of which are internal). Confidence intervals are generated using parametric bootstrap with 1,000 replications.

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.

Appendix

Polychoric Correlation

The most commonly used correlation measures 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 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

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

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

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.

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

$$P^*_{ijk}(\boldsymbol{\theta}) = P\big(Y_{ij} \geq k \mid a_i, \boldsymbol{b_i};\ \theta_j\big) = \frac{\exp\left(a_i(\theta - b_{ik})\right)}{1+\exp\left(a_i(\theta - b_{ik})\right)} \qquad (6),$$

where $P^*_{ijk}(\boldsymbol{\theta})$ is the probability of examinee j with proficiency θ_j scoring at least k on item i. This is a a_i is the discrimination parameter of item i, and b_i is the difficulty parameter of item i. Then the probability of each score is

$$P_{ik}(\theta) = P^*_{ik}(\theta) - P^*_{ik+1}(\theta). \qquad (7)$$

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

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

where $\phi(\cdot)$ is the standard normal density. Jointly considering all 6 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.

Supplementary Tables

Table A1. Inter-rater reliability with and without relationship controls

	Relationship	Percentage of Total Variability		Total	Inter-Rater Reliability		
	Controls?	Applicant	Residual	Variability	IRR Est.	LCI	UCI
Factor	Yes	30%	70%	0.87	0.3	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 (2), controlling for rater internal status and relationship effects.

Table A2. 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 (5), 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 A3. Inter-rater reliability by applicant type: internal versus external

		Applicant Type		Percentage of Total Variability		Total Variability	Inter-Rater Reliability			Novice - Experienced		
				Applicant	Residual		IRR					
		b	SE(b)				Est.	LCI	UCI	Dif. Est.	LCI	UCI
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			
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			
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			
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			
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			

Notes: Each outcome represents a separate regression model estimated using equation (5), 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.