



Improving Teacher Selection: The Effect of Inter-Rater Reliability in the Screening Process

Patricia Martinkova
Institute of Computer Science
Academy of Sciences of the Czech Republic

Department of Statistics, Center for Statistics and the Social
Sciences, University of Washington

Dan Goldhaber
Center for Education Data & Research
University of Washington Bothell

We acknowledge support from the Institute of Education Science's Researcher-Practitioner Grant Program (Grant #R305C130030) and from the National Center for the Analysis of Longitudinal Data in Education Research (CALDER) funded through grant #R305A060018 to the American Institutes for Research from the Institutes of Education Sciences, U.S. Department of Education. Patricia Martinkova gratefully acknowledges the support of the J. W. Fulbright Commission (<http://www.fulbright.cz/>) and support by Czech Science Foundation through grant #15-15856Y. This research has benefitted from the helpful input of Heather Hill, Jonah Rockoff, Nick Huntington-Klein, Roddy Theobald, Katarina Vlckova, Robert Costrell, Sarah Cannon, Elena Erosheva, Joris Mulder, Jeff Arnold, Adrian Raftery and ABC working group as well as attendees of CSSS seminar at University of Washington, and Mary Templeton and members of Spokane Public Schools' stakeholder group. Finally, we wish to thank Malcolm Wolf, Nick Huntington-Klein, Cyrus Grout, and Andrew Katz for data management and analysis support and to Jordan Chamberlain for editorial assistance. Any and all errors are solely the responsibility of the study's authors, and the views expressed are those of the authors and should not be attributed to their institutions, the study's funders, or the agencies supplying the data.

Suggested citation:

Martinkova, P. and Goldhaber, D. (2015). Improving Teacher Selection: The Effect of Inter-Rater Reliability in the Screening Process. CEDR Working Paper 2015-7. University of Washington, Seattle, WA.

© 2015 by Patricia Martinkova and Dan Goldhaber. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission, provided that full credit, including © notice, is given to the source

You can access other CEDR publications at <http://www.CEDR.us/publications.html>

Abstract

Inter-rater reliability, commonly assessed by intra-class correlation coefficient ICC, is an important index for describing the extent to which there is consistency amongst two or more raters in assigned measures. In organizational research, the data structure is often hierarchical and designs deviate substantially from the ideal of a balanced (fully crossed or nested) design. Also, often it is necessary to include covariates in the model, making it impossible to use traditional correlation-based or analysis of variance (ANOVA)-based methods for estimation of inter-rater reliability. We advocate the use of hierarchical (mixed effect model)-based methods, where variance components can be estimated by restricted maximum likelihood or Bayesian approaches. In this work, we use data from teacher hiring in Spokane public schools to demonstrate the usage of hierarchical (mixed effect) models to estimate inter-rater reliability and to demonstrate how reliability can be estimated with more complex data structures. We generally find low levels of inter-rater reliability, though this overall reliability varies according to whether the measure is assessed *across* or *within* schools, or even within job openings. We also find evidence that inter-rater reliability of some subcomponents of the hiring rubric varies according to the type of position to which applicants are applying, or according to whether the applicant is internal or from outside of the district. The direct effect of reliability on predictive power of the selection instrument is demonstrated and policy implications for public school hiring are discussed.

Key words: Inter-rater reliability, hierarchical models, intra-class correlation coefficient, linear mixed-effect model (LMM), teacher selection instruments, teacher hiring, consistency measures

1. Introduction

Each year, school districts across the United States make hundreds of thousands of teacher hiring decisions. The process of teacher selection plays a key role in influencing student outcomes. Not only is there a considerable amount of empirical evidence that the differences between teachers have consequential academic and labor market implications for students, but also much of what makes teachers effective or ineffective is associated with the individuals when they first show up at a district's doorstep to teach.¹ This alone makes the up-front choices that districts make about teacher applicants important. In addition, the hiring process itself can be costly (Barnes, Crowe and Schaefer, 2007) and it is often very difficult to remove ineffective teachers once tenured (Treu, 2014). There is a growing body of economic research pointing to the importance of the hiring process as a key workforce management tool for advancing the productivity of organizations (DeArmond et al., 2009). However, we know relatively little about the process that results in the selection of the nation's roughly 3 million public school teachers.

An unknown number of school districts utilize commercially available applicant screening and interview instruments when making hiring decisions (Young and Elli, 2002). While some of these tools have been in use for decades, their effectiveness has not been the subject of much rigorous independent analysis. A meta-analysis of the Teach Perceiver system (which was created in the 1970s) by Metzger and Wu (2008) finds only one validation study in a peer-reviewed journal. This contrasts with a large body of research that examines the statistical properties of instruments used to evaluate in-service teachers (e.g. Hill et. al, 2012, Lockwood et al, 2015).

Reliability, also referred as *generalizability* in educational literature (for extensive overview see Brennan, 2001) or *repeatability* in biological literature (see Nakagawa and Schielzeth, 2010) is a critical psychometric property on which the quality of scores is judged (AERA, APA & NCME 2014). The *inter-rater* reliability (or more precisely, its common measure, the intra-class correlation coefficient) is an important concept to assess the degree to which different raters using an evaluation instrument are likely to generate similar scores when evaluating the same subjects using the same information. It expresses the ratio of the total variation that is reproducible among repeated measurements of the same applicant. Low inter-rater reliability is an indication of imprecision in rating applicants and results in limitations in the power of the selection instrument to predict the quality of prospective teachers.

In this paper we assess the inter-rater reliability of a school-level teacher selection tool used by Spokane Public Schools (SPS) to demonstrate methods that can be applied in any admission or hiring process that employs structured ratings by multiple raters.

¹ For evidence on the importance of teacher quality see, for instance: Goldhaber et al. (1999); Nye et al. (2004), Rivkin et al. (2005); and Chetty, Friedman, & Rockoff (2014). Changing the performance of in-service teachers through policy has proved to be challenging, see for instance, Goldhaber (2015) for a review.

The mean and range of the ratings received by applicants who were rated by multiple raters on the SPS teacher selection instrument, are displayed visually in Figure 1. Each vertical line in this plot connects multiple ratings to a single applicant received during the 2009–2013 hiring period. Solid dots represent mean ratings for each applicant, and are understood as proxies for applicants' true scores, i.e. the true quality of each applicant. The vertical heterogeneity apparent in the figure indicates that some applicants were assigned very different scores at different occasions or by different raters: there are differences of up to 25 points (out of 54) in scores assigned to some applicants. Inter-rater reliability measures these discrepancies by estimating the ratio of the variance of the true scores to the total variance of ratings.

[Figure 1 about here]

There are a variety of potential explanations for inconsistencies in ratings of single applicant. For instance, it is possible the rubrics are used in different ways at different schools, or that applicants are better matched to the needs of some schools or even specific job openings. Indeed, when displaying the multiple ratings for each applicant *at a given school*, the magnitude of the inconsistencies appears to be lower.

[Figure 2 about here]

It is also possible that an applicant's quality changes over time, that raters use the rubric inconsistently, or that there is a high portion of random noise in ratings.

It is important to understand the factors that influence the inter-rater reliability. For instance, is the reliability of the ratings lower for some school or job types? Is the reliability different for internal (to Spokane) and external job applicants? Answering questions like these are important because of the impact of reliability on the predictive validity of the selection instruments. And the answers might suggest, for instance, the need to develop rating rubrics or guidelines that are school or job specific (as opposed the single rubric currently used across the district).

This study is unique in several ways. First, to the best of our knowledge, this is the first reliability analysis of an applicant screening instrument used to assess the written materials that applicants submit (as opposed to, for instance, the reliability of an in-person interview screening protocol). Second, our reliability estimates are based on observational data in an authentic setting, rather than designed as experimental studies with pre-selected raters where each rater grades all subjects. Such experimental studies allow for usage of correlation-based or ANOVA-based estimation approaches, however they might overestimate the actual reliability e.g. if the system relies on principals as raters and makes it nearly impossible to exclude individuals who are using the observational instrument in unintended or inconsistent ways (see Hill et. al, 2012). Third, as a consequence we propose to use hierarchical (mixed-effect) models for assessing

reliability. We further use parametric bootstrapping to calculate confidence intervals for inter-rater reliability estimates and we use model selection using Bayesian Information Criteria (BIC, Schwartz, 1978) and likelihood-ratio tests to answer above mentioned specific questions about reliability.

We find, in the simplest model where we look at the overall reliability across different schools and job openings, that the inter-rater reliability is only 0.27 (i.e., only 27% of the total variability in the screening scores can be attributed to the applicant quality), implying that much of the variability in applicants' screening scores is associated with the rating process itself rather than the variability in applicants' quality. Using the hierarchical models we show that measures of applicant qualities appear to be school- and job- specific (i.e. there is significant interaction between applicant and school and even between applicant and job-opening), yielding increased reliability estimates of 0.44 (0.54, respectively). These reliabilities are still considerably lower than recommendations of reliabilities in the 0.80 to 0.90 range for high-stakes decisions (Webb et al., 2006).

Using parametric bootstrapping we demonstrate varying reliability for the subcomponents of the screening rubric with some subcomponents (*Certificate & Education, Cultural Competencies*) having significantly lower inter-rater reliability than other subcomponents, suggesting that changes to specific elements of the instrument could substantially increase the overall reliability of the summative ratings.

We show that the inter-rater reliability of the screening rubric also varies according to job and location *type*, suggesting that some subcomponents of the screening rubric are less reliable for middle- and high-school positions than for elementary schools. We also observe lower inter-rater reliability (more inconsistencies between raters) in some subcomponents when used for applicants outside the district.

Finally, we show that improving the teacher selection process by increasing the inter-rater reliability would be a productive avenue of reform given evidence that the instruments, even with low reliability, appear to predict student achievement and teacher attrition (Goldhaber et al., 2014).

The paper proceeds as follows: Section 2 provides background on hiring and reliability estimates. Section 3 describes the hiring process in SPS and the data we utilize for the study. In Section 4 we describe the methodology employed. The results are presented in Section 5, and Section 6 offers policy implications and concluding thoughts.

2. Background

The potential for increasing workforce quality through more effective hiring practices, and the use of applicant selection instruments in particular, is broadly supported by research from the field of personnel economics (Heneman and Judge 2003; Shaw and Lazear 2007) and organizational psychology research (Schmitt, 2014). There are relatively few high-quality studies connecting hiring instruments to measures of workforce productivity in the case of teaching, but

the existing literature focusing on teacher selection suggest this is a promising avenue for increasing teacher quality.²

Studies of hiring outside the teaching profession suggest that the reliability of interviews can be reasonably high, in the “acceptable” range³ of 0.70. A meta-analysis of 82 studies by Conway, Jako, and Goodman (1995), for instance, finds a mean inter-rater reliability of 0.70, but there is also a good degree of heterogeneity as the 90% confidence interval for the mean ranged from 0.39 to 1.00. Not surprisingly, reliability estimates were found to be strongly related to whether candidates were interviewed by a panel or independently, the standardization of interview questions, interviewer training, and the standardization of response evaluation or methods of combining ratings.

One potential way of improving the selection process is by using *selection instruments*. Jonsson and Svingby (2007) argue that the use of rubrics can enhance the reliability of scoring and facilitate valid judgment, especially if they are analytic, topic-specific, and complemented with exemplars and/or rater training. An unknown, but not insignificant number of school districts utilize self-developed or commercially available teacher selection instruments, such as the STAR Teacher Selection Interview and instruments developed by Gallup (e.g. Metzger and Wu 2008).⁴ We could not find any independent published research on the reliability of these types of teacher selection instruments, but Metzger and Wu report that Gallup trains administrators in the use of their instrument and only those who demonstrate an 85% inter-rater reliability in item-by-item scoring with Gallup trainers are approved to use it, and the *Gallup Teacher Perceiver Interview Manual* reports a reliability of 0.76.

The selection instrument from Spokane that we focus on is different from the instruments described above because the above studies are based on in-person interviews, rather than assessments of materials found in the applications of prospective teachers. As such, it is not clear that there is any benchmark for the magnitude of the reliabilities that we might expect.⁵ Importantly, however, the quality of the applicant screening instruments used by SPS may be of greater importance than the quality of selection amongst those applicants who make it to the interview stage of the process since a relatively low percentage (27%) of applicants actually make it that far through the entire applicant review process.

Another important distinction of our study is that we are focusing on an instrument in an authentic setting. There are several downsides to estimating the reliability of selection instruments (or any instrument for that matter) that are actively being used. Most importantly,

² For examples of this work, see Dobbie (2011); Rockoff et al. (2001); and Goldhaber et al. (2014).

³ Nunnally (1979) suggested that in early stage instruments with reliability ≥ 0.70 may suffice, but in many applied research reliability of 0.80 is not nearly high enough, see also Lance et al. (2006) for discussion on this cutting value.

⁴ Metzger and Wu report that over 2,000 school districts use some type of structured selection instrument.

⁵ There is far more evidence about the inter-rater reliability of classroom observations. Hill et al (2012) found that with only one rater rating one lesson, reliabilities are often lower than 0.50, and three lessons and two raters were needed to achieve acceptable reliability of 0.70 in MQI. Similarly, Ho and Kane (2013) identified different scenarios (all with at least four observations) that produced reliability of ≥ 0.65 .

measurement designs underlying ratings gathered in organizational settings often deviate substantially from fully crossed and nested measurement designs (Hallgreen, 2012). Given that our analysis is based on nonexperimental data, there is also concern that some assumptions of the model, such as equal precision or independence of the raters may be violated. We address these issues more thoroughly in Section 6. However, having said this, estimation of reliability in authentic settings should provide better description of how well selection instruments function when utilized by practitioners, providing a better sense of their value in authentic school settings.

3. The Hiring Process in Spokane Public Schools and Data

3.1 The hiring process in Spokane Public Schools

In this study, we use the example of Spokane Public Schools (SPS) to assess the degree to which improving rater reliability can be used to improve the screening process. During the 2008-09 through 2012-13 hiring years, SPS received applications from 2,669 applicants for 526 positions that were filled via the district's standard hiring process. As described in more detail in Goldhaber et al. (2014), the hiring process follows three stages (see also Figure 3):

- 1) Intake of applications and the first stage screening⁶ of applicants by Human Resources (HR) hiring officials
- 2) Screening of applicants by school-level officials
- 3) In-person interview and hiring decision

[Figure 3 about here]

Applications are submitted through an online information system. In the application, the applicant is asked to provide contact information, information on education (e.g. college, degree, major, dates attended, additional trainings, classes or workshops), certificates and endorsements, and work experience (teaching and non-teaching). The applicant also submits supporting documentation, including a resume, cover letter, at least three letters of recommendation, certificates and licenses, and narrative statements addressing question prompts.

During the first stage of the screening, the central HR office uses a screening instrument to evaluate applicants based on the submitted materials. If applicants make it through the initial

⁶ During the period when data were collected, the district utilized a 21-point rubric. The first stage screening score is not associated with an application to any particular job and most of the first stage screenings are classified as "initial" or "general screening". An applicant is screened on this stage only the first time he or she applies for an SPS position or when new qualifications or letters of recommendations have been obtained. The 21-point rubric consisted of three criteria: "*Experience related to position*", "*Depth of skills*", and "*Quality of recommendations*".

central office screen, applicants move on to the second stage of the screening and are evaluated at the school level.

In this paper we analyze the inter-rater reliability of the second stage of the hiring process. In this stage, a principal requests a list of applicants from HR for consideration, usually specifying a cut-off score on the first stage screening rubric. Thus, the scoring in the second stage is based on a restricted range given the first stage cut (Goldhaber et al., 2014). The screening is led by the school principal and is based on the same applicant materials that are used in the first stage screening.⁷

During most of the period when data were collected (2009–2013), Spokane utilized a 54-point rubric with 9 evaluation criteria (each scored from 1–6)⁸. The components on the rubric are described in more detail in Table 1.

[Table 1 about here]

Each applicant may have been rated by one or several screeners, at the discretion of the hiring official leading the process. In the case where applicants were screened by multiple raters for the same job opening, the (average) screening scores are used to select which applicants to interview.

3.2 Descriptive portrait of applicant screenings and the raters who make them

Altogether, 2,669 applicants were screened at the district level during the 2008-09 through 2012-13 hiring years. Of the applicants screened at the first stage, 1,177 (44%) were advanced to the second stage of screening, 709 (27%) made it to the interview stage, and 374 (14%) applicants were hired.⁹ As is clear from these numbers, the two stages of the screening process play a very important role in determining which applicants are eventually hired since 73% of applicants do not advance to interviews.

⁷ The final stage of the hiring process involves an in-person interview. In contrast to screening, there are no evaluation criteria established by HR for the in-person interview. The school principal (or other hiring official) has discretion over the content—what questions are asked, how many people are on the hiring team, and how the interview is structured. Once the applicant is selected, copies of the interview questions and scoring sheets are submitted to the HR department, and after background checks the SPS makes a job offer to the top applicant.

⁸ For a short period of time, Spokane district utilized a tenth criterion *Letters of Recommendation* to rate the “quality and recentness of the recommendation, as well as the author of the letter”. Given the small amount of data, we are not analyzing the inter-rater reliability of this subcomponent and we do not include the rating in the total summative rating. However, high values of pairwise correlations with subcomponents describing applicant’s skills (see Goldhaber (2014)) suggest that letters of recommendation are main source of information on these subcomponents.

⁹ The same applicant could have been hired to different schools. For example some applicants who were hired into one school in the early years of the study data could later be rated again and hired into a different school.

We do not assess the reliability of the first stage (central office) screening process despite the large number of applicants (56%) who did not make it through this stage of the process. This is due to the fact that multiple first stage screenings generally occur when applicants' submitted materials are updated (e.g. when a new letter of recommendation is received) so we would not be able to distinguish differences in ratings related to having different raters from differences in ratings that are a result in changes to what the raters see about the applicants.

The data set consists of screenings of 3,474 individual screenings with known rater name and job location (out of original 3,955 individual screenings). Many applicants were screened multiple times—either at different schools, or at the same school for different job openings, or even for the same job opening by a different or the same rater. The *applications* data represent 1090 *applicants* rated by 137 raters¹⁰ for 484 classroom teaching job postings of 15 classroom job types¹¹ at 54 job locations in the Spokane School District during the school years ending between 2009 and 2013.

Descriptive statistics for scores on each criterion on the 54-point screening instrument, and the total summative ratings are presented in Table 2. To achieve comparable summative ratings, we adjust the sums of the subcomponents to 0–54 scale even in cases when ratings for some of the subcomponents are missing. For most of the subcomponents this is happening in less than 1% of cases, for *Certificate and Education* in less than 5% of cases and only for *Preferred Qualifications* in 26% of cases. Conversations with the district, as well as the fact that average summative 21-point scores are not statistically different between those with observed *Preferred Qualifications* subcomponent and those without, suggest that these scores are missing by accident (this last subcomponent often appeared on the second page of printed rubric and was forgotten) rather than conditional on their value, supporting the missing-at-random assumption. Results are similar if teachers with missing observations are instead dropped. We use only non-missing values when analyzing the inter-rater reliability of the individual subcomponents.

[Table 2 about here]

While many applicants were screened multiple times and many raters rated multiple times, the number of ratings for each applicant differed (see Figure A2) and also the number of ratings each rater made during the studied period 2009–2013 differed remarkably (see Figure A3).

A closer description of applicants in different steps of the admission process can be found in Goldhaber et al. (2014).

¹⁰Most of the ratings were performed by school principals or directors (68%), assistant principals (20%), coordinators (5%), or teachers (4%).

¹¹The most common classroom job types were grade teacher and specialized education teacher for elementary schools and English, math, science, social studies and special education teacher for middle and high schools. These are further discussed in Subsection 5.2

4. Methods: Hierarchical Models for Assessing Reliability

4.1 Reliability in the context of screening applicants

From an analysis of variance perspective, the estimation of inter-rater reliability is based on decomposition of variance (see also Shrout and Fleiss 1979, Webb, Shavelson and Haertel, 2006, Brennan 2001). In the simplest situation, the observed score for applicant i assigned by rater j on occasion k , Y_{ijk} , can be divided into four components

$$Y_{ijk} = \mu + A_i + B_j + e_{ijk} \quad (1).$$

The first component is the grand mean μ , constant for all applicants. The second component is applicant i 's true score A_i . The third term is the effect of rater j , denoted by B_j . And the last term is the residual e_{ijk} , which reflects the departure of observed scores on k th rating of applicant i by rater j from what would be expected given the grand mean, person's true score and effect of the rater, including a possible interaction between applicant and rater.

The variance in ratings can be decomposed into a component σ_A^2 for the systematic error-free variance among scores, a component σ_B^2 for the extent to which the scores assigned by different raters vary, and a component σ_e^2 for the residual variance (sometimes denoted as $\sigma_{AB,e}^2$ since the residual variance incorporates also the $A \times B$ interaction in addition to random error, although we cannot disentangle the two unless there are more ratings of the same applicant by the same rater):

$$\sigma_{Y_{ijk}}^2 = \sigma_A^2 + \sigma_B^2 + \sigma_e^2.$$

The **inter-rater reliability**¹² is then defined as the proportion of observed-score variance $\sigma_{Y_{pr}}^2$ due to the applicant's true-score variance σ_p^2 :

$$R = \frac{\sigma_A^2}{\sigma_{Y_{ijk}}^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_e^2}.$$

Let's assume that there is some variability between applicants. If each applicant always gets the same score (e.g. there is no variability between raters and no error), the inter-rater reliability would be equal to one: 100% of the total variability is attributable to the qualities of applicants. If on the other hand applicants receive different scores either due to the raters they receive (i.e. σ_B^2 is large), or for other random reasons (i.e. σ_e^2 is large), the inter-rater reliability will be lower.

Variance component estimators ($\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$, $\hat{\sigma}_e^2$) used to estimate inter-rater reliability can be calculated using the (ANOVA-based) moment method or using the maximum-likelihood

¹²Analogously, the **intra-rater reliability** (not a focus of this paper) is defined as proportion of the true-score variance to the total variance of ratings within single rater, in this case $R_{\text{intra}} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}$.

procedures. In this study, we suggest using restricted maximum likelihood as implemented in the lmer function of the lme4 library (Bates et al., 2014) in R (R-Core Team, 2014).¹³

We use parametric bootstrapping as implemented in the bootMer function of the lme4 library to obtain confidence intervals for mixed-effect-model-based inter-rater reliability estimates. Parametric bootstrapping is done by fitting the model and repeatedly generating data from the distribution defined by the estimated parameters. For each simulated dataset, the variance components are estimated using REML and inter-rater reliability estimate is calculated. We use $N = 1,000$ simulations to obtain a distribution of inter-rater reliabilities. We use 2.5th and 97.5th quantile of the distribution to get the 95% bootstrapped confidence interval.

4.2 Hierarchical models for testing hypotheses about reliability in teacher hiring

Model (1) presented above is appropriate if we assume that ratings should be comparable regardless of the school or job opening for which an applicant is applying or the timing of the application. In Spokane hiring the argument for this model is, for instance, that the rating process is “fair” across schools in the sense that applicants should not receive higher or lower ratings according to the particular schools in which they apply. This would also be important if Spokane seeks to use ratings derived at one school to inform hiring decisions at a different school; for instance if a school did not wish to go through an entire screening process and opted to interview top rated non-hired teachers from other schools (to our knowledge this does not happen, but it is something that Spokane has considered).

School-level reliability:

But, on the other hand, we might expect at least some subcomponents to differ across schools (and hence the summative rating) given that they focus on topics related to an applicant’s fit in a school. Also, the raters within schools might be more cohesive and there might be a shared understanding about how to interpret the rubric. Additionally, as we discuss more extensively below, a consistent rating standard across schools could inhibit schools from differentiating the applications they receive since not all schools appear to have a distribution of applicants that are equally qualified.

To acknowledge that applicants may be differently suited to teach at different schools (e.g. schools with a more diverse student body might, for instance, have different expectations about cultural competency than schools with a high percentage of white students), we need to include in the model the interaction term AS_{il} which accounts for potential applicant-school matching effects:

$$Y_{ijkl} = \mu + A_i + B_j + AS_{il} + e_{ijk} \quad (2).$$

¹³ See the Appendix B for selected R code.

With model (2), when estimating reliability, we are looking at how consistent the ratings are *if the application was rated for the same school*:

$$R = \frac{\sigma_A^2}{\sigma_{Y_{ijkl}}^2} = \frac{\sigma_A^2 + \sigma_{AS}^2}{\sigma_A^2 + \sigma_B^2 + \sigma_{AS}^2 + \sigma_e^2}.$$

Job-level reliability:

Similarly, to acknowledge that applicants might be differentially suited for different individual job openings within a school (e.g. due to the composition of students in the class for which the teacher is hired), we can look at the consistency of ratings assigned only for the same job opening.¹⁴

For this purpose, we need to include the interaction term AJ_{im} which accounts for potential applicant-job matching effects. $Y_{ijklm} = \mu + A_i + B_j + AS_{il} + AJ_{im} + e_{ijklm}$ (3).

With model (3), when estimating reliability, we are looking at how consistent the ratings are *if the application was rated for the same job opening (i.e. also for the same school)*:

$$R = \frac{\sigma_A^2}{\sigma_{Y_{ijklm}}^2} = \frac{\sigma_A^2 + \sigma_{AS}^2 + \sigma_{AJ}^2}{\sigma_A^2 + \sigma_B^2 + \sigma_{AS}^2 + \sigma_{AJ}^2 + \sigma_e^2}.$$

Including other effects is possible as well. For example, we tried to also include the school effect (with and without the interaction term—models 2a, 2b), the job effect (models 3a, 3b), year effect, or effect of the prescreening date (which may be an indicator for change of the application materials). Mentioned effects were added subsequently and we used the Bayesian Information Criterion (BIC, Schwartz, 1978) for model selection as well as other criteria such as the Akaike Information Criterion (AIC), and the likelihood-ratio test (see Table 4).

4.3 Reliability differing by applicant type or job category:

We further want to argue that inter-rater reliability varies with applicant type: we expect that it is higher for ratings of internal applicants than for applicants from outside of the district. Our assumption is based on the premise that there is higher rater variability as well as higher residual error in ratings of external applicants.

¹⁴ Different job categories can be open at the same school, but *job opening* specifies both job category and job location. Ratings for job openings with unknown job location were removed from analysis.

Analogously, we want to argue that the inter-rater reliability varies by job location or job category: we expect that it is higher for elementary schools for which the rubric was primarily developed.

To get the description of the reliabilities in the groups, we first fit the job-level model (3) separately on the data of the individual groups (e.g. separately on internal and external applicants), to obtain estimates of inter-rater reliability for the two groups. We also build bootstrapped confidence intervals separately using two fitted models for the two groups (see Figures 5 and Table A2 for internal vs. external applicants, and Figures 6 and 7 as well as Table A3 for Elementary vs. MS/HS positions).

To test the significance of varying reliability based on group, we need to include group effect into the model, as well as to allow the variances of the random effect components to vary based on the group. We build the model in several steps. We start with job level model (3b) which was selected as the best fitting model (see Table 4):

$$Y_{ijklm} = \mu + A_i + B_j + AS_{il} + J_m + AJ_{im} + e_{ijklm} \quad (3b)$$

First we add only the fixed group effect

$$Y_{ijklm} = \mu + \omega_i \beta_{A1} + A_i + B_j + AS_{il} + J_m + AJ_{im} + e_{ijklm} \quad (4a)$$

where $\omega_i = 1$ if i is an internal applicant and equals 0 otherwise and where β_{A1} is a difference in fixed effects for internal and external applicants. Next, we allow the applicant variance to vary for the two groups (interval vs. external applicants):

$$Y_{ijklm} = \mu + \omega_i \beta_{A1} + \omega_i A_{0i} + (1 - \omega_i) A_{1i} + B_j + AS_{il} + J_m + AJ_{im} + e_{ijklm} \quad (4b)$$

where applicant random effect is allowed to have different variance for external and internal applicants: $A_{0i} \sim N(0, \sigma_{0A}^2)$ and $A_{1i} \sim N(0, \sigma_{1A}^2)$.

In further steps, we can similarly also allow the variance of the rater effect (model 4c) and error term (model 4e) to vary across the two groups¹⁵.

Each model (4a–4e) is fitted on the whole dataset and compared with other models using the BIC or likelihood-ratio test. Models (4a–4e) allow for simultaneous estimation of the inter-rater reliability for each group, as well as for bootstrapped confidence intervals of the estimates.

4.4 Improving reliability

¹⁵ It was not possible to fit the model with varying variances of interactions (model 4d) due to low number of observations.

Given the potential for concerns about low reliability, a natural question relates to how reliability can be increased. Possible approaches are providing more exact criteria for evaluation of the applicant materials, training the raters, or, if ratings are mostly dependent on recommendation letters, ensuring confidentiality of the recommendation letters and providing the evaluation criteria to those who write them. Given the differences in inter-rater reliability of the subcomponents, another promising avenue seems to be maximizing reliability by component weighting, which we do not explore in this paper. .

Averaging ratings from higher number of raters

Yet one way to increase the reliability of the measure is to use the average score from J raters within job opening

$$\bar{Y}_{i.klm} = \frac{\sum_{j=1}^J Y_{ijklm}}{J}.$$

Assuming model (3) and assuming that the raters are equally precise, such an average will have smaller variance

$$\sigma_{\bar{Y}_{i.klm}}^2 = \sigma_A^2 + \sigma_B^2/J + \sigma_{AS}^2 + \sigma_{AJ}^2 + \sigma_e^2/J,$$

and higher inter-rater reliability

$$R_k = \frac{\sigma_A^2 + \sigma_{AS}^2 + \sigma_{AJ}^2}{\sigma_{\bar{Y}_{i.klm}}^2} = \frac{\sigma_A^2 + \sigma_{AS}^2 + \sigma_{AJ}^2}{\sigma_A^2 + \sigma_B^2/J + \sigma_{AS}^2 + \sigma_{AJ}^2 + \sigma_e^2/J} \quad (5).$$

4.5 The impact of inter-rater reliability on predictive validity

Efforts to increase the inter-rater reliability are important especially given the fact that measurement error (i.e. low reliability) implies weaker correlations between variables. Assume two variables measuring some underlying properties with error

$$Y = \mu + \varepsilon_\mu, \quad Z = \theta + \varepsilon_\theta,$$

e.g. screening score Y measuring applicant's quality μ at the time of hiring, and Z measures θ , the teacher's value added to student's achievement. Assuming the errors are uncorrelated with each other and with the measurements Y and Z , the correlation can be expressed as a product of the correlation of the underlying properties and their reliabilities R_μ, R_θ :

$$\text{corr}(Y, Z) = \frac{\text{cov}(\mu + \varepsilon_\mu, \theta + \varepsilon_\theta)}{\sqrt{\text{var}(\mu + \varepsilon_\mu) \text{var}(\theta + \varepsilon_\theta)}} = \text{corr}(\mu, \theta) \sqrt{R_\mu} \sqrt{R_\theta} \quad (6)$$

In this work, to calculate the attenuated correlation $\text{corr}(Y, Z)$ between screening scores and value added (see column 4 of Table 5), we have to accommodate for the fact that for some value

added scores of given applicant/year combination we have multiple screening scores from the given hiring year (preceding the year the value added is counted from). We therefore calculate the correlation several times ($N=1,000$). In each loop we randomly pick one of the screening scores for a given individual and hiring year to use it in computation of the correlation coefficient.

Estimates of reliability can then be used to estimate the correlation of the variables in the case where there is less or no measurement error (both random error and the systematic measurement error caused by differences between raters). The correlation between the underlying properties in our case gives the predictive validity of the teacher quality measured at the time of hiring for teacher effectiveness in the sense of value added.¹⁶

5. Results

In this section, we begin by assessing the inter-rater reliability estimates of the summative ratings (Section 5.1, Table 3). In Section 5.2 we focus on the inter-rater reliability of different subcomponents of the instrument (Figure 4, Table 4 and Table A1). In Section 5.3 we describe differences in the inter-rater reliability for different groups: we will focus on the differences in inter-rater reliability between internal and external applicants (Figure 5, Table A2), elementary vs. middle/high school job openings (Figure 6, Table A3), and individual job categories (Figure 7). In Section 5.4, we focus on the effects of increasing reliability of the subcomponents by averaging scores over two or three raters, and finally, we estimate the effect of increasing reliability on the ability of the screening scores to predict teacher value added (Table 5).

5.1 Estimates of inter-rater reliability of summative ratings

As reported in Table 3, only 27% of total variability can be attributed to differences between applicants. This is displayed visually in Figure 1, which plots multiple ratings for each application on a given data aggregation level.

At the school level (model 2), reliability estimates are nearly twice as high (0.44, see Table 3, row 2). However, there are still cases where there is considerable heterogeneity in the ratings at this level (see Figure 2), which amounts to differences of more than 20 points on the rubric (over 2 standard deviations) for some applications.

As expected, the reliability within job openings (model 3) is a bit higher (0.54), but the increase that we see associated with moving from the school to the job level is far smaller than the increase associated with moving from the applicant to the school level. In other words, the aspects of application packages that are viewed differently are more associated with school level differences in ratings than with different positions within the school. This could be a result of applicants having qualifications that make them more well-suited to one school versus another,

¹⁶ In regression analysis, this in effect biases the regression coefficients toward zero, also called attenuation (see Green, 2000, p. 377).

which might imply that the between school differences in ratings are driven by particular subcomponents on the instrument. We discuss this possibility more extensively below. However, it also might reflect differences in rating standards across schools.

[Table 3 about here]

The Bayesian Information Criteria (BIC) as well as likelihood-ratio test (not displayed) suggests that the best fitting model out of models (1–3) is model 3, providing evidence for the varying applicant effect at different schools and even job openings.

5.2 Reliability of screening score subcomponents

Looking at individual subcomponents (see Figure 4, and in more detail in Table A1) the observed pattern is similar to the pattern observed for the total score: In all subcomponents there are significant applicant-school matching effects. The fact that subcomponent ratings of individual applicants vary across schools is understandable given the heterogeneity in students across schools in Spokane.¹⁷ However another plausible explanation as we noted above is that there is a more shared understanding amongst raters of what they are looking at when evaluating applicants within schools.

Comparison of the models based on BIC also suggests that there are significant applicant-job matching effects (as implied in model 3) in almost all components, except for *Training*, *Instructional Skills* and *Cultural Competencies*, where applicant effect varies only on school level (model 2). As can be expected, remarkable differences in inter-rater reliability between the two models (2) and (3) are in mostly job-related subcomponent *Preferred Qualifications* (Figure 4).

Looking at within school estimates exclusively (blue vertical lines in Figure 4), the inter-rater reliability varies widely across the subcomponents. The inter-rater reliabilities on the *Certificate & Education* or *Cultural Competency* components, for instance, are quite low relative to the estimates of inter-rater reliability on *Training*, *Experience*, or *Instructional Skills*.

A possible explanation is that the underlying traits of applicants that the subcomponents are designed to measure may not be terribly clear to raters. For instance, low estimates of reliability in *Cultural Competency* and *Interpersonal Skills* correspond with stakeholders' anecdotes suggesting that often they are not sure what score to assign for this category, or that recommendation letters are silent on this category. Those who write the recommendation letters

¹⁷For instance there are some elementary schools where 25% of students are underrepresented minorities whereas other schools have less than 3% of their students in these subgroups.

may not know that they should include information relevant to these criteria or they don't know what to write¹⁸.

On the other hand, the subcomponent *Instructional Skills* has higher inter-rater reliability (0.46 at the school level) than the summative rating (0.44). Relatively high inter-rater reliability is also achieved for other subcomponents describing applicant skills: *Experience*, *Classroom Management*, and *Flexibility*.

[Figure 4 about here]

5.3 Reliability differing by applicant type or job category:

Internal and external applicants

Focusing on applicants who previously taught in Spokane or who completed their teacher training in Spokane, we observe that for some subcomponents (*Management*), there is significantly higher inter-rater reliability than in external applicant (see Figure 5, Table A2). We also observe a clear pattern that the inter-rater reliability is higher for internal applicants across other subcomponents as well.

This might be caused by the fact that raters are more familiar with the schools (in Spokane) where applicants received their training (e.g. the *Training* subcomponent has a reliability of 0.50 on the job level for internal applicants, but only 0.39 for external) so know better how to interpret information in an application. Another possibility is that those writing the recommendation letters for internal applicants tend to be from Spokane so are likely to be more familiar with the rating system and needs of the district, and so speak directly to these in their letters.¹⁹

[Figure 5 about here]

Using more complex models 4a – 4e which include applicant type (internal/external) and fitting them on the whole dataset allows for a direct test of the applicant-type effect on overall rating (model 4a) and on variance components (models 4b-4e) and thus on inter-rater reliability.

For each subcomponent, the models including applicant type fit better (as indicated by lowest BIC, see Table 4). The optimal model for each subcomponent includes at least fixed effect of applicant type (Internal vs. External), suggesting that on average the internal applicants get

¹⁸ The estimates of *intra*-rater reliability (not shown in this paper) were relatively low, reinforcing the idea that raters do not always have a very concrete idea of how to rate application materials

¹⁹ It is also possible that there are informal feedback mechanisms for internal applicants. While this is not what raters in the district report, they may be in contact with those writing letters within the district.

higher ratings than external applicants. For *Certificate & Education, Training and Management*, the model allowing for different variability of applicant effect for internal and external applicants is selected as the optimal model, thus implying different inter-rater reliability estimates for internal and external applicants. For *Training*, the optimal model also allows variability of rater effect to be different for internal and external applicants.

[Table 4 about here]

Reliability for Different Job Locations and Different Job Categories

In Figure 6 (and in more detail in Table A3) we display the inter-rater reliability estimates calculated separately for elementary and for middle/high school job openings. Similarly, in Figure 7 are displayed estimates for different job categories, keeping in mind that for each job category we analyze a different sample of applicants, and that for some categories the sample size is low.²⁰

Two patterns are notable. First, the reliability for the screening of special education teachers is higher than for other positions in many subcomponents, a not so surprising finding if one assumes that there is a high degree of consensus amongst raters about the skills needed to be effective with special education students. However, the reliability at the elementary level is higher than at the middle and high school levels. This is surprising given the above reasoning and the fact that middle/high school positions are more specialized. One potential explanation for the lower reliability at the middle and high school levels is that the raters differ in terms of their level of specialization. For instance, most ratings include a principal, but we might expect other members of the team reviewing applicants at the middle school and, especially, the high school level to also include subject specialists.

[Figure 6 about here]

[Figure 7 about here]

5.4 Estimates of the effects of number of raters on reliability and predicted validity

As can be seen from equation (5) above, one way to increase reliability is by increasing the number of raters, and by making hiring decisions based on the average of their ratings. Having the estimates of proportions of variability, we can further estimate the way in which reliability differs in the presence of multiple raters as a prediction of how the reliability of the overall estimate might be improved if multiple raters were used more widely. We use reliability

²⁰We focus on the 7 most common job categories based on the number of applications.

estimates from Table A1 (model 3) as well as estimates of variability components and formula (5) to estimate reliability in the case of two and three raters (see first three columns of Table 5).

To illustrate the impact of low/high reliability in one of the measures of validity, given the hiring instrument and its components, we estimated the correlations between the 54-point screening ratings (and its components, respectively) and teacher value-added in mathematics, using the data from Goldhaber et al. (2014). To assess the predictive validity of the rubric and its components in the case of one rating, we used applicants with available value added estimates and screening scores from the previous hiring year. Estimated correlations can be seen in the fourth column of Table 4.²¹ Using equation (6), we evaluate correlation in the case where there is no error in the ratings (last column of Table 5). Estimates of reliability (first three columns of Table 5) were then used to calculate the expected correlation scores for average ratings of one, two and three raters.²²

[Table 5 about here]

To increase reliability to be ≥ 0.65 , two raters would be enough for the summative rating and some subcomponents (*Experience*, *Instructional Skills*, *Preferred Qualifications*). The use of three raters is estimated to increase reliability all the way to 0.80 on these subcomponents (see first three columns of Table 5). On the other hand, some subcomponents are less affected by the number of raters. For instance, even with three raters, the reliability of some subcomponents still doesn't reach a value of 0.65 (*Certificate & Education*, *Cultural Competency*). Cases like these suggest that the subcomponent is used inconsistently or that the underlying quality the subcomponent is trying to assess is not well defined or understood by the raters.

Two factors explain the correlation between subcomponent ratings and value added scores (see Table 5, columns 4–7): the relationship between the quality being measured by a subcomponent and teachers' value added and the level of reliability of the subcomponent rating (or the value added estimate). Not surprisingly, the subcomponents with low inter-rater reliability correlate weakly with the teacher value added estimates. On the other hand, while the *Instructional Skills* subcomponent reached the highest inter-rater reliability, *Classroom Management*, which has lower reliability, predicted teacher value added much better. Besides these two components, two other subcomponents correlated strongly with subsequent value added: *Training* and *Flexibility*.

²¹ These differ from the estimates of Goldhaber et al. (2014) since these are bivariate correlation coefficients, and Goldhaber et al. report multivariate regression coefficients.

²² Specifically, we use the estimates of correlation from the bootstrap estimate in the fourth column of Table 8 as an estimate of $Corr(Y, Z)$ in equation (4). Then, we take reliability estimates from the first three columns of Table 8, and alter those estimates given the model of how reliability changes with the addition of more raters in equation (3). Those adjusted estimates give the fifth through last columns in Table 8.

For these subcomponents, increasing the reliability leads to large increases in the estimated predictive power.

6. Discussion, Policy Implications, and Conclusions

In this study we examine the reliability of an instrument designed to assess the written materials that teacher applicants submit when applying for teaching positions. We find that reliability, whether it is measured across schools, within schools, within jobs in schools, or within raters is low, at least in terms of the standards (0.8 to 0.9) generally recommended for making high stakes decisions (Webb et al., 2006). And this screening instrument is indeed used for very high-stakes decisions as it determines whether teacher applicants advance through the application pipeline in Spokane. However, we have no real baseline for comparison as this is to our knowledge the first study that focuses on the reliability of an instrument used for screening application materials in this context.

It is also worth mentioning several limitations and caveats associated with this work. First, the reliability is estimated in an authentic setting. This may yield a good estimate of the likely reliabilities we would see were the same screening instrument to be used in other school districts, (as opposed to, for instance, an analysis that is based on judgments made by specially trained raters). But, in the analysis process we have made various assumptions—such as the independence of the raters—that are not directly verifiable. We also face a situation where only a subset of applicants is rated by multiple raters, raising issues of generalizability. For the findings to be generalized to the whole sample, the decision to rate multiple times cannot be correlated with the quality of the ratings or with the distribution of the applicants' true scores. Were it the case, for instance, that schools that tend to screen each applicant multiple times did so as a consequence of having a very collegial teaching staff or a close working relationship between principals, assistant principals, and teachers (all of whom rate some applicants), one might also anticipate that they would tend more generally to agree with one another. In such cases, the ratings from different raters may be more similar to each other than they would be from schools who did not adopt multiple screenings, i.e. these schools might also have high inter-rater reliability due to unobserved factors unrelated to the number of ratings. On the other hand, the existence of multiple raters might reflect a desire to get a broader perspective about applicants to mitigate the possibility of, for instance, a lack of trust in a principal's judgment alone. In this scenario the ratings might be expected to be more heterogeneous and our estimate of inter-rater reliability could underestimate the true reliability.

Not surprisingly our estimates show that the inter-rater reliability increases significantly in all subcomponents moving from comparisons of applicants between and across schools to just within schools. This might suggest that schools are differentially rating applicants in part because they have different needs, i.e. an applicant with the same background legitimately receives differential ratings at different schools given her/his suitability to teach in a particular context.

Another simple explanation for the low level of reliability across schools is that schools have different standards when they are judging candidates. If so, applicants do not have an equal

probability of advancing through the Spokane pipeline regardless of the school to which they apply, but it is also not clear that this would be desirable from the district's perspective. The reason is that schools attract different applicant pools. For instance, some Spokane schools tend to screen a much stronger pool of applicants (as measured by their licensure test scores or first stage screening ratings). With the current 1–6 rating scale and common rating standards across schools, the raters would have difficulties discriminating amongst these stronger applicants as they would be forced to assign them primarily fives and sixes on the subcomponents to be consistent with the ratings that applicants at other schools receive. One potential solution to this type of dilemma is to allow schools to assign a decimal point to the 1–6 rating scale, allowing schools to be consistent across buildings and differentiate applicants who apply to a particular school (at the decimal place).

Yet even when we move to an assessment of within school and within job ratings, the reliabilities are still relatively low. This might be expected given that many of the criteria upon which applicants are rated are based on subjective materials (such as recommendation letters) submitted by applicants. The results suggest that raters may not always have a very concrete idea of how to rate application materials. This, combined with the importance of this screening instrument for making teacher hiring decisions, implies that the district could benefit from more training on the rating instrument, particularly on the subcomponents that have low reliabilities.

We have demonstrated two approaches to address the question whether the inter-rater reliability differs for two groups (e.g. internal and external applicants). First approach used job-level model applied separately on two datasets – dataset of internal and dataset of external applicants (Figure 5 and Table A2). Bootstrapped 95% confidence intervals suggest that there is significantly lower inter-rater reliability in ratings of external applicants in some subcomponents (e.g. *Management*) and a clear pattern for all other subcomponents.

The second approach utilized more complex models (4a – 4e) with random slopes which allowed for different mean ratings and different random effect variances for two groups. These models fit better for all subcomponents (see Table 4), suggesting significant differences in variance components for internal and external applicants, thus in inter-rater reliability.

However, we acknowledge that the two approaches might lead to different conclusions. There are tradeoffs that need to be considered when choosing one of the approaches. The more complex model takes into account all data and especially dependencies that may be present in ratings of internal and external applicants when rated by the same rater. However, significant differences in variance components in the two groups may lead to only minor and insignificant differences in inter-rater reliabilities. This is partly due to the fact that inter-rater reliability is defined as a ratio of variances, therefore even if one variance component differs between internal and external applicants, this still does not have to lead to a difference in inter-rater reliabilities if the difference is compensated by a difference in another variance component.

There are several extensions of the analyses that might be fruitful. First, we used linear-mixed effect models even for subcomponents which are rated on a 0–6 scale. However, with ordinal rating scales, the variability of the scores is lower when the ratings are closer to the boundary

values 0 or 6. A possible solution could be to use ordinal models to define inter-rater reliability. Also, we have so far not included in the analysis the fact that the subcomponents are correlated. Finally, the restricted-likelihood methods used for estimation could further be compared with Bayesian approaches, which might be more appropriate given the sparse nature of the ratings.

Thus far we have focused on the shortcomings of the instrument used by Spokane. This should not mask the fact that, despite seemingly low levels of reliability, the instrument has been shown to have predictive validity in terms of teachers' future value added and probability of attrition (Goldhaber et al., 2014). And, as we illustrate, for some of the subcomponents the level of predictive validity could be increased substantially by adding raters. We cannot say for certain whether an investment in training or extra raters would be cost-effective. However, given the importance of teacher quality in influencing student outcomes, investing resources in improving the rating system would be a promising avenue to pursue.

REFERENCES

- AERA, APA, NCME (2014). *Standards for Educational and Psychological Testing*.
- Barnes, G., Crowe, E., & Schaefer, B. (2007). The Cost of Teacher Turnover in Five School Districts: A Pilot Study. National Commission on Teaching and America's Future.
- Bates, D., Maechler M., Bolker B. and Walker S (2014). lme4: Linear mixed-effect models using Eigen and S4. R package version 1.1-7. URL <http://CRAN.R-project.org/package=lme4>
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-added and Student Outcomes in Adulthood. *American Economic Review*. 24(9), 2593-2632.
- Conway, JM., Jako, R and Goodman, D. (1995). A Meta-Analysis of Interrater and Internal Consistency Reliability of Selection Interviews. *Journal of Applied Psychology* 80(5):565–79. Retrieved (<http://doi.apa.org/getdoi.cfm?doi=10.1037/0021-9010.80.5.565>).
- DeArmond, M, Shaw, K., and Wright. P. (2009). Zooming In and Zooming Out: Rethinking School District Human Resource Management. In *Creating a New Teaching Profession*, edited by Dan Goldhaber and Jane Hannaway. Urban Institute Press.
- Dobbie, W. (2011). *Teacher Characteristics and Student Achievement: Evidence from Teach For America*. Harvard University. Cambridge, MA.
- Goldhaber, D., Brewer, D., and Anderson, D. (1999). A Three-way Error Components Analysis of Educational Productivity. *Education Economics*, 7(3), 199-208.
- Goldhaber, D. (2006). "Are teachers unions good for students?" In *Collective Bargaining in Education: Negotiating Change in Today's Schools*, edited by Jane Hannaway and Andrew Rotherham. Cambridge, MA: Harvard Education Press, pp. 141-157.
- Goldhaber, D., Grout, C., Huntington-Klein, N. (2014) Screen Twice, Cut Once: Assessing the Predictive Validity of Teacher Selection Tools. CEDR Working Paper, WP 2014-09.
- Goldhaber, D. (2015) Teachers Clearly Matter, But Finding Effective Teacher Policies Has Proven Challenging. In *Handbook of Research in Education Finance and Policy*, 2nd edition. Edited by Helen F. Ladd, Margaret E. Goertz. Routledge, New York.
- Greene W.H. (2000). *Econometric Analysis*, 4th edition. Prentice-Hall, New Jersey.
- Hallgreen, K.A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol.* 8(1):23-34.
- Hannaway, J. and Rotherham, A. (2006). *Collective Bargaining in Education: Negotiating Change in Today's Schools*. Cambridge, MA: Harvard Education Press.

- Harris, D., Rutledge, S., Ingle, W., and Thompson, C. (2010). Mix and Match: What Principals Really Look for When Hiring Teachers. *Education Finance and Policy* 5(2):228–46. Retrieved (<http://www.mitpressjournals.org/doi/abs/10.1162/edfp.2010.5.2.5205>).
- Heneman, HG. and Judge, T. (2003). *Staffing Organizations*. 4th ed. Middleton, WI: McGraw-Hill/Mendota House.
- Hess, F.M. and Kelly, A. (2006). Scapegoat, albatross, or what? The status quo in teacher collective bargaining. In *Collective Bargaining In Education: Negotiating Today's Schools* (ed. J. Hannaway and A. Rotherham), 53-87.
- Hill HC, Charalambous CY, Kraft MA. (2012). When rater reliability is not enough: Teacher observation systems and a case for the G-study. *Educational Researcher*. 41(2):56-64.
- Ho A.D., Kane T.J. (2013). *The Reliability of Classroom Observations by School Personnel*. Retrieved from: http://www.metproject.org/downloads/MET_Reliability_of_Classroom_Observations_Research_Paper.pdf
- Jonsson, A and Svingby, G. (2007). The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences. *Educational Research Review* 2(2):130–44. Retrieved June 05, 2014 (<http://linkinghub.elsevier.com/retrieve/pii/S1747938X07000188>).
- Lance C.E., Butts M. M., Michels L. C. (2006). The sources of four commonly reported cutoff criteria what did they really say? *Organizational Research Methods* 9 (2), 202-220.
- Metzger, S. and Wu, MJ. (2008). Commercial Teacher Selection Instruments: The Validity of Selecting Teachers Through Beliefs, Attitudes, and Values. *Review of Educational Research* 78(4):921–40. Retrieved June 13, 2014 (<http://rer.sagepub.com/cgi/doi/10.3102/0034654308323035>).
- Nakagawa, S. and Schielzeth H. (2010): Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews*, 85, 935-956.
- NCTQ (2014). Teacher Contract Database. National Council on Teacher Quality. URL <http://www.nctq.org/districtPolicy/contractDatabaseLanding.do>.
- Nye, B., Konstantopoulos, S, & Hedges, L.V. (2004). How Large are Teacher Effects? *Educational Evaluation and Policy Analysis*, 26, 237-257.
- Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.
- Putka, D. and Hoffman, B. (2013). Clarifying the Contribution of Assessee-, Dimension-, Exercise-, and Assessor-Related Effects to Reliable and Unreliable Variance in Assessment Center Ratings. *The Journal of Applied Psychology* 98(1):114–33. Retrieved December 30, 2014 (<http://www.ncbi.nlm.nih.gov/pubmed/23244226>).

- R Core Team (2014). "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>
- Riley, P., Fusano, R., Munk, L., and Peterson, R. (2002). Contract for failure: The impact of teacher union contracts on the quality of California schools. Pacific Research.
- Rivkin et al. (2005) Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2), 417–458.
- Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy*, 6(1), 43–74.
- Schmitt, N., editor (2014). *The Oxford Handbook of Personnel Assessment and Selection*. Oxford Library of Psychology.
- Schwartz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, Vol. 6(2), 461-464.
- Shaw, K and Lazear, E. (2007). Personnel Economics: The Economist's View of Human Resources. *Journal of Economic Perspectives* 21(4):91–114.
- Shrout, P and Fleiss, J. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin* 86(2):420–28.
- Strunk, K. (2011). Are Teachers' Unions Really to Blame? Collective Bargaining Agreements and Their Relationships with District Resource Allocation and Student Performance in California. *Education Finance and Policy*, 6(3), pp. 354-398.
- Strunk, K. and Grissom, J. (2010). Do strong unions shape district policies? Collective bargaining, teacher contract restrictiveness, and the political power of teachers' unions. *Educational Evaluation and Policy Analysis*, 32(3), 389-40.
- Treu, R. M. Vergara vs. State of California Tentative Decision (2014).
- Webb, M.W., Shavelson, R.J, and Haertel, E.H. (2006). Reliability Coefficients and Generalizability Theory. In *Handbook of Statistics*, Vol. 26, Elsevier B.V.
- Young, I. Phillip and Dane A. Delli. (2002). The Validity of the Teacher Perceiver Interview for Predicting Performance of Classroom Teachers. *Educational Administration Quarterly* 38(5):586–612.

Table 1: Screening criteria of the 54-point screening rubric

Criterion	Look for...
Certificate and Education	Note completion of course of study; certificate held (current or pending); education
Training	Look for quality, depth and level of candidate's additional training related to position
Experience	Note degree to which experience supports the prediction of success, not just the number of years. A beginning candidate could be rate highly
Classroom Management	Look for specific references to successful strategies. This may not mean quiet and orderly, but planned and directed. Effectively handles large/small or ethnically/sociologically diverse groups; develops routines and procedures to increase learning, establishes clear parameters, and responds appropriately.
Flexibility	Note multiple endorsements, activity, coaching interests, student, building or district, or community support. Willing to learn new concepts and procedures, successfully teachers a variety of assignments, effectively uses various teaching styles.
Instructional Skills	Look for specific references in support of skill in this area – plans, implements, evaluates, relates to students, creative, multiple approaches, monitors and adjusts, uses culturally responsive strategies appropriate to age, background and intended learning of students.
Interpersonal Skills	Develops and maintains effective working relationships with diverse staff, students, parents/guardians, and community.
Cultural Competency	Look for specific references to successful strategies for building and maintaining a relationship with each student and their family. This may not be explicitly mentioned, but the following strategies offer some evidence of cultural competency: specific instructional strategies providing each student access to a rigorous curriculum, inclusive/respectful language about students and families, a belief that all children can achieve at high levels, mention of conflict resolution/restorative practices, specific instructional strategies for integrating culturally responsive materials which are also rigorous, and appropriate statements about their work with diverse populations. Note relevant training, course work, authors/book titles listed.
Preferred Qualifications	Look for possession of qualifications as indicated on job posting.

Table 2: Summary statistics

	Obs.	Mean	SD	Min.	Max.
Summative Rating	3 474	37.88	7.84	8	54
Certificate and Education	3 326	5.05	1.09	0	6
Training	3 451	3.99	1.33	0	6
Experience	3 473	4.06	1.20	0	6
Management	3 446	4.08	1.15	0	6
Flexibility	3 459	4.24	1.11	0	6
Instructional	3 463	4.18	1.17	0	6
Interpersonal	3 453	4.37	1.09	0	6
Cultural	3 443	3.99	1.18	0	6
Preferred Qualifications*	2 560	3.89	1.53	0	6

Notes: * Item “Preferred Qualifications” was often forgotten as it appeared on the other side of the printed form.

Table 3: Estimates of the variance components and inter-rater reliability for summative ratings

	Percentage of total variability					Total variability	BIC	Inter-rater reliability		
	Applicant	Rater	Appl:School	Appl:Job	Residual			Estimate	LCI	UCI
model 1	27%	22%	-	-	51%	64.65	23228.50	0.27	0.23	0.30
model 2	20%	22%	25%	-	34%	63.27	23104.48	0.44	0.40	0.49
model 3	20%	21%	18%	17%	25%	62.83	23076.20	0.54	0.49	0.59

Notes: LCI – lower bound of bootstrapped 95% confidence interval, UCI – upper bound of bootstrapped 95% confidence interval

Table 4: Model comparison using BIC

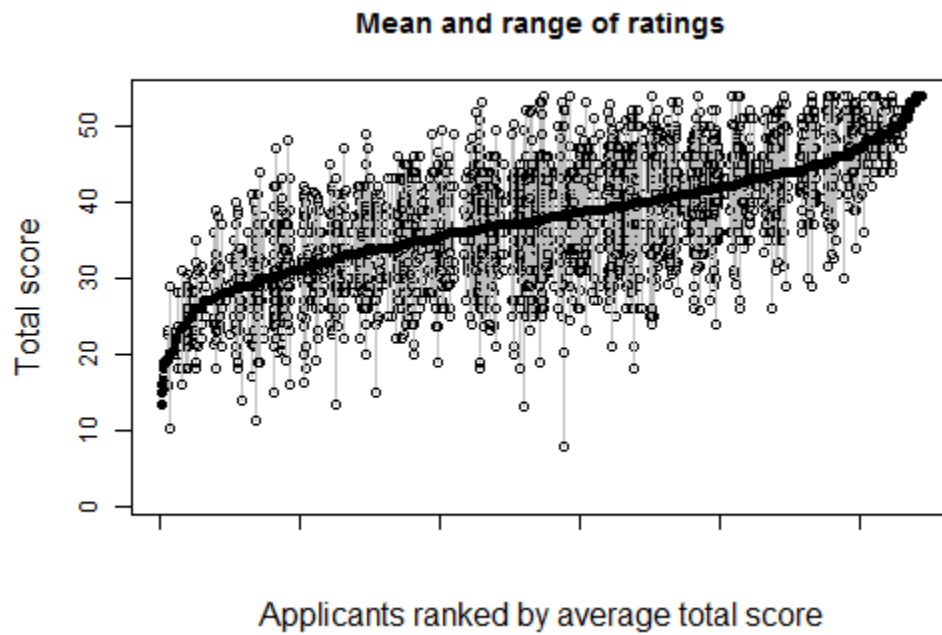
		Summative rating	Subcomponents								
Model	Model description		Certif & Ed	Training	Experience	Mangmnt.	Flexibility	Instruct.	Interpers.	Cultural	Pref. Qual.
model 1	Applicant and Rater effect only	23228.5	8539.8	11074.7	10585.9	10263.9	9998.4	10295.1	9764.5	10394.9	9096.0
model 2	+ Applicant:School interaction	23104.5	8451.4	11007.8	10495.1	10208.6	9924.7	10205.6	9714.5	10350.6	8991.1
model 3	+ Applicant:Job interaction	23076.2	8434.6	11009.7	10490.9	10207.0	9920.4	10207.8	9710.5	10355.1	8955.3
model 3b	+ Job effect	22956.8	8035.2	10927.2	10441.4	10124.0	9840.5	10177.2	9632.8	10249.1	8619.3
model 4a	+ Applicant type (Internal vs. External)	22867.0	8027.8	10875.5	10415.4	10072.4	9793.4	10111.2	9597.1	10212.5	8584.7
model 4b	+ Applicant effect allowed to differ for Internal/External	22870.7	7986.1	10887.7	10431.1	10068.3	9794.9	10124.5	9612.8	10216.8	8587.6
model 4c	+ Rater effect allowed to differ for Internal/External	22876.8	7997.7	10870.8	10439.9	10077.3	9809.1	10139.0	9625.3	10230.6	8596.9
model 4e	+ Residual allowed to differ for Internal/External	22884.0	7982.2	10878.9	10448.1	10079.1	9815.9	10136.6	9631.4	10233.8	8604.3

Notes: Lowest BIC for each subcomponent typed in bold, indicating optimal model from the set of models.

Table 5: Increasing the reliability and predictive validity by using average of more raters

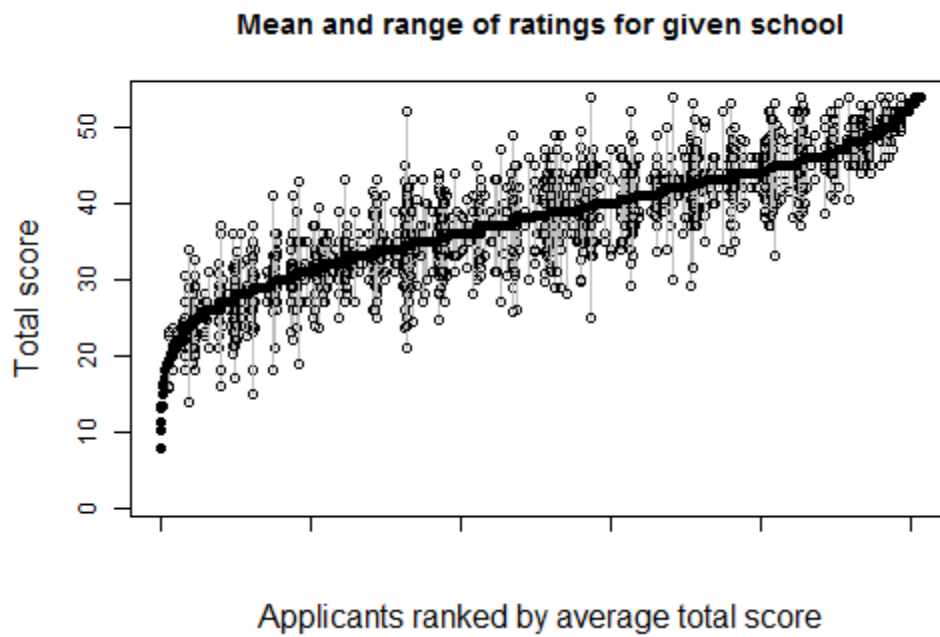
	Estimated reliability for average of:			Estimated correlation with VA for average of:			
	1 rater	2 raters	3 raters	1 rater	2 raters	3 raters	Infinite # of raters
Summative rating	0.54	0.70	0.78	0.16	0.18	0.19	0.22
Certificate and							
Education	0.33	0.50	0.60	0.03	0.04	0.04	0.05
Training	0.48	0.64	0.73	0.18	0.21	0.22	0.26
Experience	0.51	0.68	0.76	0.11	0.13	0.13	0.15
Management	0.43	0.60	0.69	0.32	0.38	0.41	0.49
Flexibility	0.43	0.60	0.69	0.20	0.24	0.25	0.30
Instructional	0.50	0.67	0.75	0.15	0.17	0.18	0.21
Interpersonal	0.43	0.61	0.70	0.07	0.08	0.09	0.11
Cultural	0.36	0.53	0.63	0.04	0.05	0.05	0.07
Preferred							
Qualifications	0.60	0.75	0.82	0.07	0.08	0.08	0.09

Figure 1: Mean and range of summative ratings of applicants rated multiple times



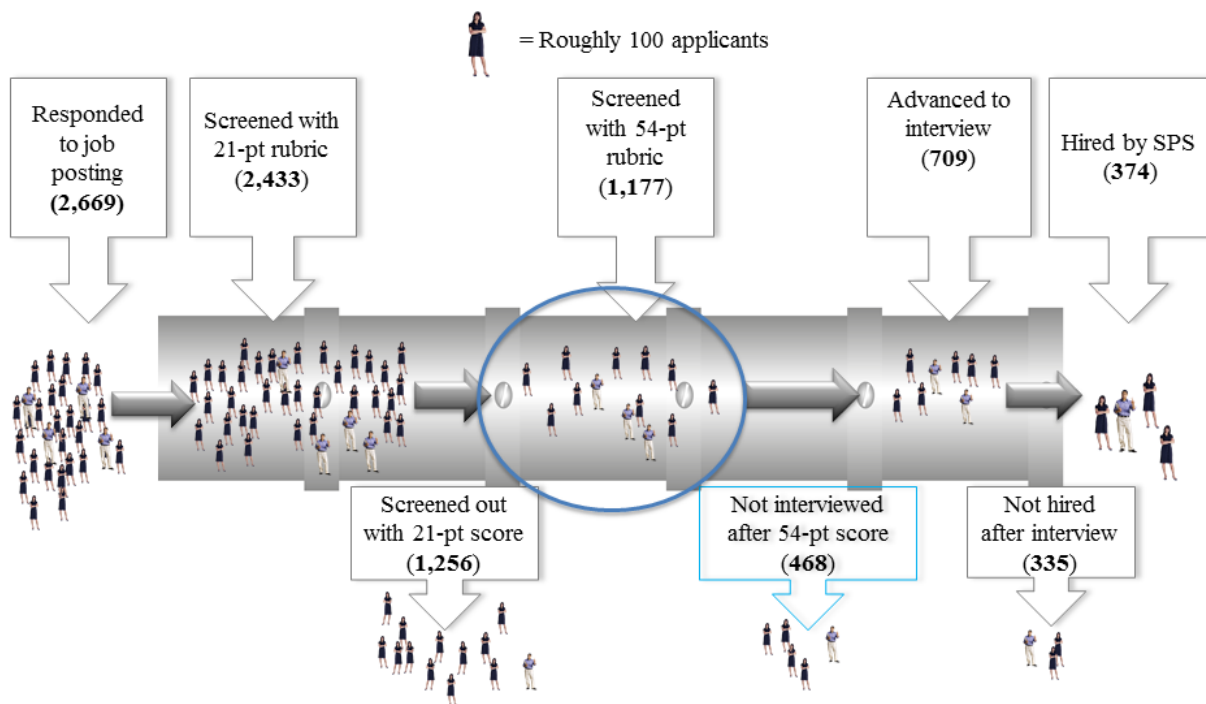
Notes: Each vertical line connects ratings given to single applicant during period 2009–20 (the same or at a different school).

Figure 2: Mean and range of summative ratings of applicants rated multiple times at the same school



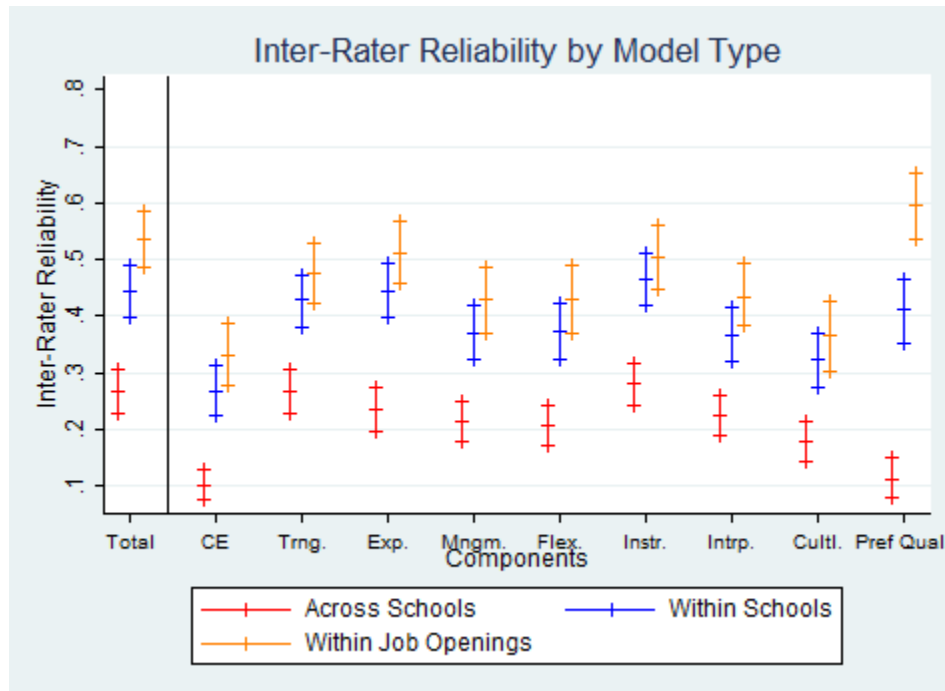
Notes: Each vertical line connects multiple ratings given to single applicant **at the same school** during period 2009–2013.

Figure 3: Hiring process in Spokane public schools



Notes: Numbers indicate total number of applicants between years 2008/09 and 2012/13. Blue circle represents population used for estimation of reliability of 54-point rubric.

Figure 4: Estimates of inter-rater reliability for the summative rating and for subcomponents



Notes: Estimates and bootstrapped 95% confidence intervals for three models: Across Schools (model (1)), Within Schools (model (2)) and Within Job Openings (model (3)). Comparison of the three models (using BIC) suggests that applicant effect varies for schools for Summative rating as well as for all subcomponents, and that applicant effect varies also for Job openings for total score as well as for most of the subcomponents (all, except for *Training*, *Instructional Skills* and *Cultural Competencies*), suggesting that the estimate of inter-rater reliability based on Model 3-Within Job Opening) is the most appropriate one. For further details, see Table A1.

Figure 5: Estimates of inter-rater reliability for internal and external applicants

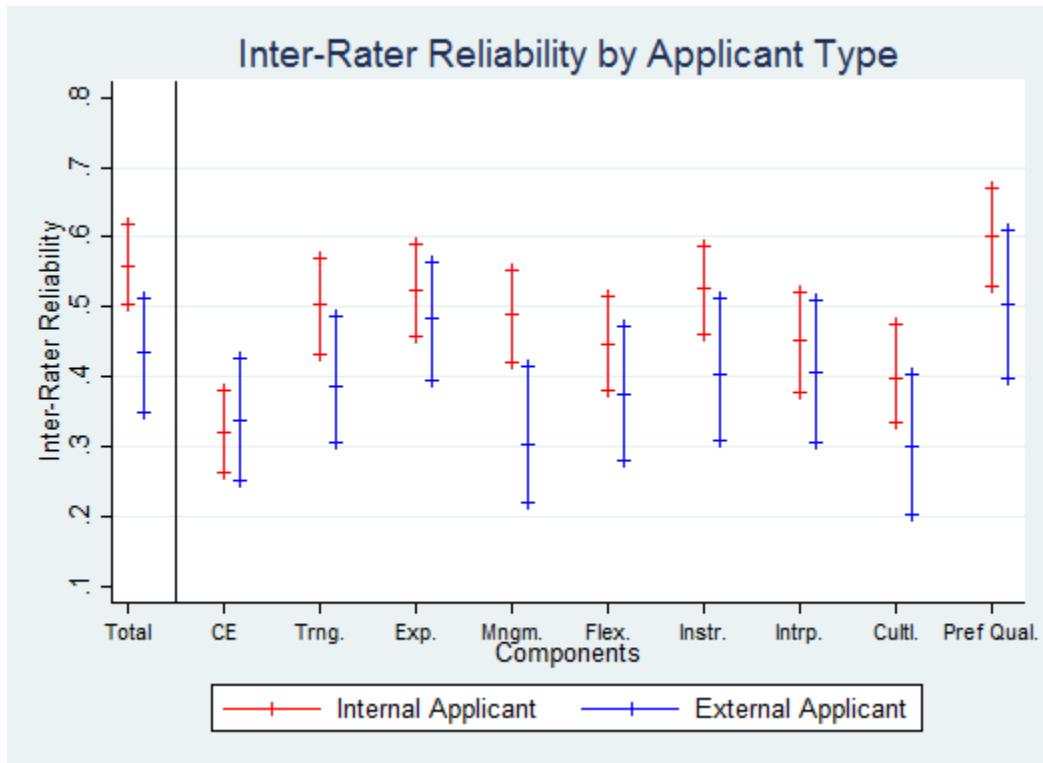
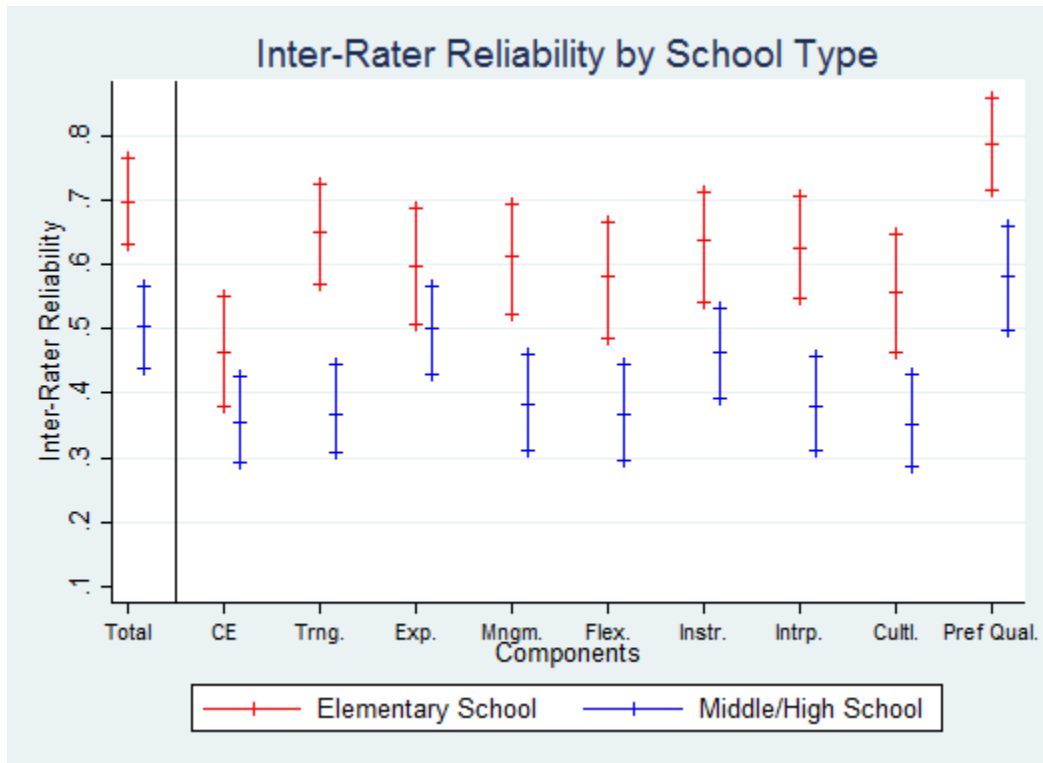


Figure 6: Estimates of inter-rater reliability by school type



Appendix A – Supplemental Tables and Figures

Table A1: Variance decomposition and estimates of inter-rater reliability for models (1 – 3)

Model (1) – Across Schools	Percentage of total variability					Total variability	BIC	Inter-rater reliability		
	Applicant	Rater	Appl:School	Appl:Job	Residual			Estimate	LCI	UCI
Summative Rating	27%	22%	-	-	51%	64.65	23228	0.27	0.23	0.30
Certificate and Education	10%	43%	-	-	47%	1.25	8540	0.10	0.08	0.13
Training	27%	13%	-	-	60%	1.77	11075	0.27	0.23	0.30
Experience	23%	13%	-	-	64%	1.46	10586	0.23	0.20	0.27
Management	21%	13%	-	-	65%	1.36	10264	0.21	0.18	0.25
Flexibility	21%	17%	-	-	63%	1.29	9998	0.21	0.17	0.24
Instructional	28%	12%	-	-	60%	1.40	10295	0.28	0.24	0.32
Interpersonal	22%	18%	-	-	60%	1.24	9765	0.22	0.19	0.26
Cultural	18%	18%	-	-	64%	1.46	10395	0.18	0.14	0.21
Preferred Qualifications	11%	21%	-	-	68%	2.44	9096	0.11	0.08	0.15
Model (2) – Within Schools	Percentage of total variability					Total variability	BIC	Inter-rater reliability		
	Applicant	Rater	Appl:School	Appl:Job	Residual			Estimate	LCI	UCI
Summative Rating	20%	22%	25%	-	34%	63.27	23104	0.44	0.40	0.49
Certificate and Education	5%	42%	22%	-	31%	1.23	8451	0.27	0.23	0.31
Training	21%	13%	22%	-	44%	1.76	11008	0.43	0.38	0.47
Experience	17%	12%	27%	-	43%	1.46	10495	0.44	0.40	0.49
Management	16%	13%	21%	-	50%	1.34	10209	0.37	0.32	0.42
Flexibility	15%	16%	23%	-	46%	1.27	9925	0.37	0.32	0.42
Instructional	22%	12%	25%	-	42%	1.38	10206	0.46	0.42	0.51
Interpersonal	17%	17%	19%	-	46%	1.22	9715	0.37	0.32	0.41
Cultural	13%	18%	19%	-	50%	1.45	10351	0.32	0.28	0.37
Preferred Qualifications	5%	19%	37%	-	40%	2.39	8991	0.41	0.35	0.47

Model (3) – Within Job Openings	Percentage of total variability					Total variability	BIC	Inter-rater reliability		
	Applicant	Rater	Appl:School	Appl:Job	Residual			Estimate	LCI	UCI
Summative Rating	20%	21%	18%	17%	25%	62.83	23076	0.54	0.49	0.59
Certificate and Education	5%	42%	15%	13%	25%	1.22	8435	0.33	0.28	0.39
Training	21%	13%	18%	9%	40%	1.75	11010	0.48	0.42	0.53
Experience	17%	12%	22%	12%	37%	1.46	10491	0.51	0.46	0.57
Management	16%	13%	15%	12%	44%	1.34	10207	0.43	0.37	0.49
Flexibility	15%	16%	16%	12%	40%	1.26	9920	0.43	0.37	0.49
Instructional	21%	12%	21%	8%	38%	1.37	10208	0.50	0.45	0.56
Interpersonal	17%	17%	14%	13%	40%	1.21	9711	0.43	0.38	0.49
Cultural	13%	18%	16%	7%	46%	1.45	10355	0.36	0.30	0.43
Preferred Qualifications	5%	18%	27%	28%	22%	2.37	8955	0.60	0.54	0.65

Notes: Lowest Bayesian Information Criterion (BIC) within the three models for each subcomponent is printed in bold.

Table A2: Variance decomposition and estimates of inter-rater reliability for internal and external applicants

Model (3) for Internal Applicants		Percentage of total variability					Total variability	Inter-rater reliability		
	Applicant	Rater	Appl:School	Appl:Job	Residual		Estimate	LCI	UCI	
Internal Applicants	Summative Rating	17%	21%	24%	15%	23%	60.46	0.56	0.50	0.62
	Certificate and Education	2%	46%	15%	15%	22%	1.13	0.32	0.26	0.38
	Training	18%	12%	22%	11%	37%	1.66	0.50	0.43	0.57
	Experience	15%	11%	26%	11%	37%	1.40	0.53	0.46	0.59
	Management	15%	13%	19%	15%	38%	1.26	0.49	0.42	0.55
	Flexibility	15%	16%	20%	10%	40%	1.22	0.45	0.38	0.52
	Instructional	20%	11%	26%	7%	36%	1.28	0.53	0.46	0.59
	Interpersonal	15%	15%	18%	12%	40%	1.13	0.45	0.38	0.52
	Cultural	12%	16%	23%	5%	44%	1.37	0.40	0.34	0.47
	Preferred Qualifications	6%	19%	34%	20%	21%	2.37	0.60	0.53	0.67

Model (3) for External Applicants		Percentage of total variability					Total variability	Inter-rater reliability		
		Applicant	Rater	Appl:School	Appl:Job	Residual		Estimate	LCI	UCI
External Applicants	Summative Rating	17%	28%	11%	16%	28%	60.96	0.44	0.35	0.51
	Certificate and Education	15%	36%	6%	13%	30%	1.39	0.34	0.25	0.43
	Training	19%	23%	19%	0%	38%	1.78	0.39	0.31	0.49
	Experience	19%	16%	20%	10%	35%	1.47	0.48	0.40	0.56
	Management	15%	17%	15%	0%	52%	1.39	0.30	0.22	0.41
	Flexibility	14%	21%	5%	18%	42%	1.28	0.38	0.28	0.47
	Instructional	17%	14%	17%	6%	46%	1.41	0.40	0.31	0.51
	Interpersonal	13%	22%	10%	17%	37%	1.30	0.41	0.31	0.51
	Cultural	15%	19%	6%	10%	51%	1.49	0.30	0.20	0.40
	Preferred Qualifications	0%	25%	14%	36%	25%	2.28	0.50	0.40	0.61

Table A3: Variance decomposition and estimates of inter-rater reliability by school type.

Model (3) for Elementary Schools		Percentage of total variability					Total variability	Inter-rater reliability		
		Applicant	Rater	Appl:School	Appl:Job	Residual		Estimate	LCI	UCI
Elementary Schools	Summative Rating	23%	18%	12%	35%	13%	62.49	0.70	0.63	0.76
	Certificate and Education	4%	36%	12%	30%	17%	1.02	0.46	0.38	0.55
	Training	29%	8%	15%	21%	27%	1.80	0.65	0.57	0.73
	Experience	16%	13%	22%	22%	28%	1.49	0.60	0.51	0.69
	Management	21%	12%	7%	33%	27%	1.42	0.61	0.52	0.69
	Flexibility	16%	16%	12%	30%	25%	1.33	0.58	0.49	0.66
	Instructional	25%	10%	22%	17%	26%	1.43	0.64	0.54	0.71
	Interpersonal	16%	13%	12%	34%	25%	1.21	0.62	0.55	0.70
	Cultural	7%	21%	16%	32%	23%	1.53	0.56	0.46	0.65
	Preferred Qualifications	5%	19%	21%	52%	3%	2.36	0.79	0.72	0.86
Model (3) for Middle and High Schools		Percentage of total variability					Total variability	Inter-rater reliability		
		Applicant	Rater	Appl:School	Appl:Job	Residual		Estimate	LCI	UCI
Middle and High Schools	Summative Rating	19%	19%	22%	10%	30%	59.12	0.50	0.44	0.57
	Certificate and Education	6%	38%	22%	8%	27%	1.28	0.36	0.29	0.43
	Training	17%	16%	19%	2%	47%	1.70	0.37	0.31	0.45
	Experience	20%	9%	19%	11%	41%	1.34	0.50	0.43	0.57
	Management	13%	10%	22%	3%	52%	1.22	0.38	0.31	0.46
	Flexibility	13%	15%	21%	3%	49%	1.17	0.37	0.30	0.45
	Instructional	20%	9%	20%	6%	45%	1.28	0.46	0.39	0.53
	Interpersonal	18%	16%	15%	6%	46%	1.15	0.38	0.31	0.46
	Cultural	19%	12%	16%	1%	53%	1.35	0.35	0.29	0.43
	Preferred Qualifications	4%	16%	30%	24%	26%	2.34	0.58	0.50	0.66

Figure A1

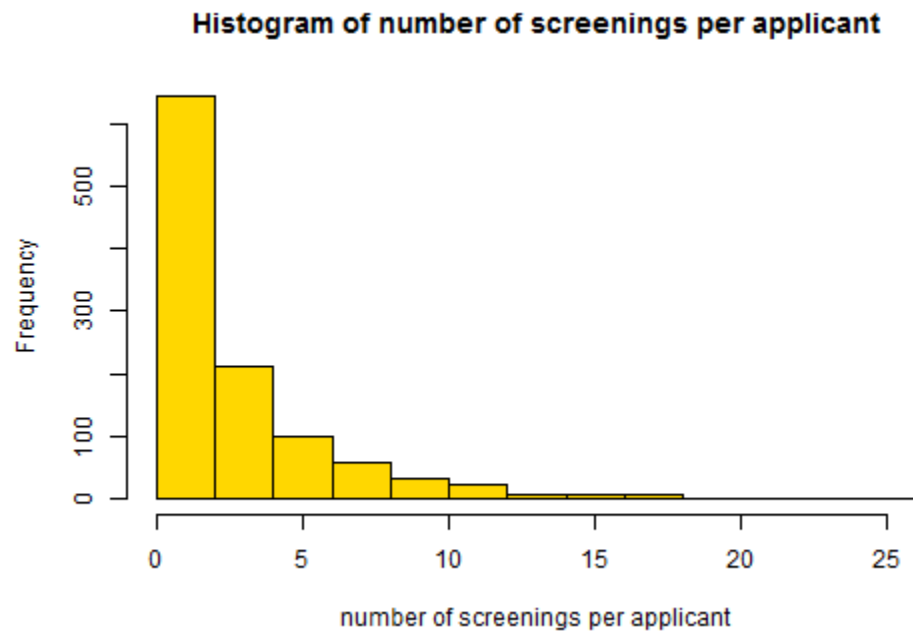
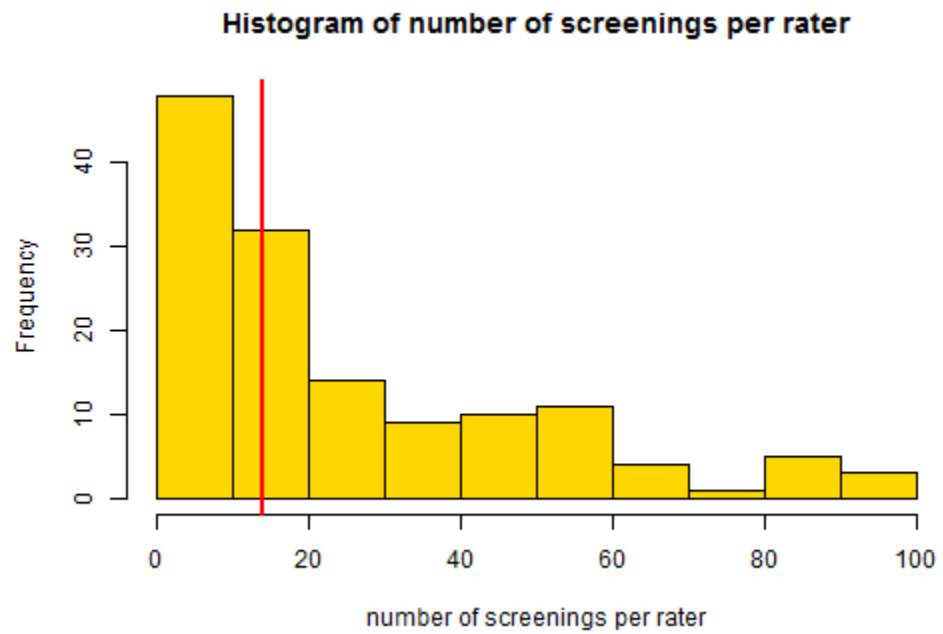


Figure A2



Appendix B – Selected R code

```
# Overall level:
model1 <- lmer(total~1+(1|applicant)+(1|rater), data=screening)

# School level:
model2a <- lmer(total~1+(1|applicant)+(1|rater)+(1|school),data=screening)
model2b <- lmer(total~1+(1|applicant)+(1|rater)+(1|school)+
(1|applicant:school), data=screening)
model2 <- lmer(total~1+(1|applicant)+(1|rater)+(1|applicant:school),
data=screening)

# Job level:
model3a <- lmer(total~1+(1|applicant)+(1|rater)+(1|applicant:school)+
(1|job_id), data=screening)
model3b <- lmer(total~1+(1|applicant)+(1|rater)+(1|applicant:school)+
(1|job_id)+ (1|applicant:job_id), data=screening)
model3 <- lmer(total~1+(1|applicant)+(1|rater)+(1|applicant:school)+
(1|applicant:job_id), data=screening)

# Models including applicant-type effect:

# model 4a: internal get higher score
model4a3b <- lmer(total~1+internal+(1|applicant)+(1|rater)+
(1|applicant:school)+(1|job_id)+(1|applicant:job_id), data=screening)

# model 4b: applicant variability varies for internal/external applicants
model4b3b <- lmer(total~1+internal+(1|rater)+(0+internal|applicant)+
(1|job_id)+(1|applicant:school)+(1|applicant:job_id), data=screening)

# model 4c: also rater variability varies for internal/external applicants
model4c3b <- lmer(total~1+internal+(0+internal|rater)+
(0+internal|applicant)+(1|applicant:school)+(1|applicant:job_id)+(1|job_id),
data=screening)

# model 4e: The error term varies for internal/external
error1 <- 1:dim(screening)[1]
error1[screening$internal] <- 0
error1f <- as.factor(error1)

model4e3b <- lmer(total~1+internal+(0+internal|rater)+(0+internal|applicant)
+(1|applicant:school)+ (1|job_id)+(1|applicant:job_id)+(1|error1f),
data=screening))
```

```

# Model comparison

BIC(model1, model2, model3b, model3, model4a3b, model4b3b, model4c3b,
model4e3b, anova(model1, model2, model3b, model4a3b, model4b3b, model4c3b,
model4e3b)
anova(model1, model2, model3b, model4a3b, model4b3b, model4c3b, model4e3b)

# Estimation of inter-rater reliability in model (3):

model <- lmer(total~1+(1|applicant)+(1|rater),data=screening)
varAppl <- as.numeric(VarCorr(model)["applicant"])
varRater <- as.numeric(VarCorr(model)["rater"])
varApplSch <- as.numeric(VarCorr(model)["applicant:school"])
varApplJob <- as.numeric(VarCorr(model)["applicant:job_id"])
varResid <- sigma(model)^2

rel1 <- (varAppl + varApplSch + varApplJob)/(varAppl + varRater + varApplSch
+ varApplJob + varResid)

# Bootstrapped 95% confidence interval for inter-rater reliability

bIRRmodel4 <- bootMer(model, function(mm) (VarCorr(mm)$applicant[,] +
as.numeric(VarCorr(mm)["applicant:school"]) +
as.numeric(VarCorr(mm)["applicant:job_id"]))/ (VarCorr(mm)$applicant[,] +
VarCorr(mm)$rater[,] + as.numeric(VarCorr(mm)["applicant:school"]) +
as.numeric(VarCorr(mm)["applicant:job_id"]) + sigma(mm)^2), 1000)

lci <- quantile(bIRRmodel2$t, 0.025)
uci <- quantile(bIRRmodel2$t, 0.975)

# Increasing reliability by averaging ratings of 2 or 3 raters:
rel2 <- (varAppl + varApplSch + varApplJob)/(varAppl + varRater/2 +
varApplSch + varApplJob + varResid/2)
rel3 <- (varAppl + varApplSch + varApplJob)/(varAppl + varRater/3 +
varApplSch + varApplJob + varResid/3)

# Increasing correlation with VAM by increasing reliability
cor1 <- c(0.16, 0.03, 0.18, 0.11, 0.32, 0.20, 0.15, 0.07, 0.04, 0.07)
corInf <- cor1/sqrt(rel1)
cor2 <- cor1/sqrt(rel1)*sqrt(rel2)
cor3 <- cor1/sqrt(rel1)*sqrt(rel3)

```