Employment Selection Instruments—What We Have Learned from Ten Years of Research

Howard Ebmeier, University of Kansas
Amy Dillon, Shawnee Mission, KS Public Schools
Jennifer Ng, University of Kansas

Over the last ten years a group of professors, graduate students, and HR professionals have been engaged in an extensive examination of employment screening instruments commonly used in schools and the development of new instruments useful for early identification of exemplary applicants in a number of employment classifications. These studies have resulted in the publication of numerous scientific journal articles, countless dissertations, and several technical papers. The following article summarizes our conclusions in a number of areas important for HR professionals. As a disclaimer, all of us are currently involved in developing employment interview instruments for AASPA. The majority of conclusions reached in this article flow directly from that work.

The Importance of Known Metrics

Because of the serious consequences of decisions made during the initial employment screening process, all instruments employed during that process are expected by the public, courts, and prospective employees to conform to certain psychometric standards. Chief among these standards are validity, reliability, job relatedness, and that the instruments have been developed and used following recommended practices suggested by the National Council for Measurement in Education and adopted by EEOC in their guidelines titled, “Uniform Employee Selection Guidelines”. Indeed, the “gold standard” is that the instrument be developed using the guidelines suggested by NCME/EEOC, be peer reviewed (typically by being published in an academic journal), and that the reliability and validity be independently confirmed by external replication studies. Most of the instruments employed by schools today for student testing or by school psychologists meet this “gold standard”.

Unfortunately, from our rather extensive examination of existing commercial instruments available to human resource management (HRM) departments for screening employees, many fall short of these standards. Very few of the commonly available instruments have been published in peer reviewed journals and replication studies by external independent reviewers are lacking. What is more troubling is when independent reviews have been conducted (often in the form of dissertations), a great number of these studies offer little support for the claims of the commercial companies.

The authors would like to acknowledge the work of the following individuals who contributed in a major way to this paper: Drs. Tim Allshouse; Jennifer Beutel; Patrick Cowan; George Crawford; David Cox; Darren Dennis; Erin Dugan; Mary Elizabeth Green; Ken Emley; Tina Hale; Dale Longenecker; Michael Reik.; Mark Schmidt; Vicki Smith; Scott Springfield; Gary Stevenson; and Michael Weishaar.


An earlier version of this paper was presented at the annual conference of AASPA in Kansas City in 2007.


See the following for critical reviews: Koerner, Robert. (2007), The Relationship Between the TeacherInsight Interview Scores and Student
Unfortunately, most commercial companies are reluctant to release reliability and validity data to external researchers claiming that the information is proprietary in nature. Transparency for prospective clients is even lacking as basic psychometric data is generally not included in promotional material and often technical manuals, common to most standard instruments, are lacking altogether.

The degree to which these issues surface varies widely across commercial companies. We would advise that before a school district adopts one of these instruments, they demand basic psychometric information from all companies under consideration. Failing to obtain this information, we would suggest that the company be removed from further consideration. In addition, it would be helpful for a competent psychometrical in the district (school psychologist for example) to review all the technical data and read a representative sample of dissertations or published articles about the instruments under consideration before a final decision is reached. It never makes sense to employ a tool without supporting validity and reliability evidence regardless of the cost, ease of administration, or reputation of the company.

Knowing What the Instrument Measures

Whether you select one of the commercial instruments or develop one yourself, be sure you know what it measures. Some instruments measure teacher personality variables (empathy, drive, openness, collegiality, support, conscientiousness, etc.) thought to distinguish effective from non-effective teachers. Other instruments measure teacher traits such as persistence, commitment to teaching, likelihood of success in urban teaching, and predicted length of teaching career. Still others focus on classroom teacher behaviors derived from the teacher effectiveness literature (discipline practices, time on task procedures, techniques for instructional delivery). Our advice would be to select instruments and questions that are most closely related to the job description and to keep them focused on having the candidate tell you what they would do or have done in a given situation. From our research, while personality questions will separate effective teachers from ineffective teachers, situational/job related questions are generally more predictive in identifying discrete differences.

Most of the instruments define teacher effectiveness in terms of supervisor ratings. While this seems reasonable, other measures such as student satisfaction, parent satisfaction, residual gain on state tests, and miscellaneous student affective outcomes are also possible.

For an overview of these studies and comparison chart see www.people.ku.edu/~howard/ICIS.html.


Given the varying definitions of effective teachers and teaching, a single instrument may not work for everyone. We would suggest that districts carefully consider what teacher behaviors and dispositions they most desire, then select or construct instruments that focus on those elements of effectiveness. This might mean that differing parts of the organization use different instruments or emphasize different scales within a singular instrument to help select the type of candidate in which they are most interested.

**Question Development**

There are two ways in which to derive the questions for use on an initial employment selection system. The first involves the generation and field testing of a large number of questions covering many aspects of the teaching profession, teacher personality, teacher traits, and classroom competencies. Responses to these questions from employees deemed outstanding and marginal on whatever dimension of interest (supervisor evaluation, test score attainment, employment longevity, etc.) are then compared statistically to determine which questions best separate the two groups. Questions that are deemed the best predictors of effectiveness or ineffectiveness are then included on the selection instrument. The second method is to use existing job descriptions or standards developed by national organizations to identify desirable behaviors for the given employee group. These standards then serve as the basis for question development with the questions subjected to expert panel review and extensive field testing to remove biased, confusing, or otherwise inappropriate questions.

While both methods of question development are satisfactory, we prefer the second for four reasons. First, constructing questions based on job descriptions and national recommendations is conceptually more appropriate than random selection of questions based on their ability to statistically separate applicants. Each question is included because of a theoretical basis apparent to national experts and directly linked to job performance. Second, from our discussions with many applicants, most prefer questions that appear to link directly to a particular job. Internet chat rooms are full of complaints about the nature of the questions of some of the commercial instruments for their lack of job relevance. While the questions might be good predictors, job applicants often have difficulty seeing how the questions relate to teaching. This lack of apparent relevance often creates negative images of the district that would “make them go through this silly game” as one recent teacher candidate related on the internet. Third, we believe it is easier to defend job related questions if necessary. Basing employment decisions on job relevant factors is the general expectation of the various courts and EEOC. Finally, using the extreme group comparison method may isolate questions that separate current good from poor teachers but remember you are attempting to predict future behavior not current differences. These different response patterns may have developed after several years of teaching and not be present in beginning teachers. If that is correct, then screening beginning teachers for behaviors that only develop after years of classroom experience would not be very productive.

Questions selected for inclusion on the instrument should meet several criteria. First they should be brief, jargon free, and only inquire about a singular idea. For example, the question, “How would you structure the curriculum and learning environment in your classroom to teach urban children?” contains multiple questions and is, therefore, difficult to succinctly answer and score. A better question might be, “What should 4th grade students know about geometry?” Second, we believe questions should be designed to elicit responses that demonstrate conceptual understanding on the part of the candidate. Understanding the reasoning ability and thought processes of a candidate we believe is very important. Simple knowledge of a given curriculum or textbook series seems less significant than understanding what children need to know at a given age level. Third, designing questions to assess how well the candidate matches the district’s philosophical orientations does not seem healthy for the organization. Schools as organizations need individuals with different belief systems to meet new challenges. For example, while empathy is a desirable characteristic of teachers, employing only individuals who score highly on an employment instrument emphasizing empathy

---

8 Actually, employing several different instruments can enhance the predictive power of the selection process. Two well designed instruments collecting different data should produce better decisions.
will, in the long run, distort the diversity of the teaching staff in undesirable ways. The employment screening instrument should help provide a snapshot of the candidate but not make the employment decision for you. Obviously, we are not supportive of arbitrary cut-off scores on any employment screening instrument. Their proper use is only for data collection.

**Number of Questions Needed**

We have spent considerable time running Monte-Carlo simulations in an attempt to answer this question. From these studies, we would estimate that 5-7 well designed questions with rubrics are adequate for each scale. Questions directly related to classroom practices are better predictors of supervisor ratings and require fewer questions to reach a higher level of validity than other types of questions. Asking more than 15 questions does not appreciably improve the predictive validity of the instrument and appears just to waste the time of the candidate and the administrator. On the other hand, an instrument reporting results for 10 separate scales would require approximately 70 questions. The more reported scales, the more questions needed.

Estimates of the quality of each scale can be obtained from the reliability estimates provided in the technical manual. Generally one should expect reliability estimates over 0.70 for each scale and 0.90 for the instrument as a whole.

We have discovered through field trials of potential interview systems the importance of allowing the interviewer the ability to skip questions if appropriate. Sometimes even the most carefully conceived questions are inappropriate for a given position. For example, asking a potential music teacher about methods of teaching reading may not be helpful. Also, sometimes a candidate simply misunderstands the nature and intent of the question and gives a nonsense answer. The ability to quickly move to an alternative question of a similar nature keeps the interview progressing smoothly and avoids the necessity of having the interviewer attempt to rephrase the question.

For on-line screening instruments, adaptive algorithms should be employed to improve the efficiency of the instrument. Asking 50 sequential and non-adaptive questions over the internet can actually result in less accurate information than employing 20 well selected questions based on the candidate’s previous responses. We are not aware of any commercial instrument that utilizes such adaptive technology, but they should definitely consider that possibility. Most commercial student testing companies have utilized such technology for years.

**Interview Scoring Rubric Development**

For face-to-face interviews, rubrics provide useful guidance for the interviewer concerning how to score the candidate’s answer. They also help establish inter-rater reliability across interviewers and help force the interviewer to make a series of independent judgments about a candidate rather than a single general impression at the end of the interview. Using rubrics, which require evaluation of the responses to individual questions, also allows differential weightings of questions and grouping of questions into multiple scales.

Our experience indicates that three scoring levels are adequate both in terms of the ease of interviewer training and reliability. Increasing the number of possible scoring alternatives above three tends to decreases scoring consistency and increase training difficulty. Including less than three scoring alternatives does not typically provide enough discrimination power.

The better the original interview question (clear conceptual focus, singular in nature, avoids jargon, etc.), generally the easier it is to construct the various levels of rubrics. The key is to anticipate how a strong, average, and weak candidate might respond to each of the prompt questions. Importantly, assessors must be able to agree on decisions about the quality of answers compared to the standards, and these decisions must be able to be defended and explained in order to provide content validity. One should be careful not to confuse the quality of an answer with the length or the number of points made unless they are valid.

Once the rubrics have been created, one should check for face and concurrent validity by using member and expert checking. By asking those “in the field” to validate the content of the rubrics, you can be more assured that the questions and rubrics are realistic. We try to construct the rubrics such that about half of the candidates’ answers are scored in the middle with twenty-five percent in the lower and upper third of the distribution.
Potential Bias

Unintentional bias is always a concern when using employment screening instruments. Although most of the commercial companies have not reported data regarding bias, from our work and the extant literature, if the interview instrument is well constructed with carefully conceived rubrics, there seems to be little gender or experience bias inherent in the instrument itself. Age bias, however, does seem to be a potential problem at least from our studies and inferentially from the literature of some of the commercial companies. For example, one of the companies cautions readers that their instrument might not be appropriate for individuals under the age of 30. Our informal conversations with many teachers and administrators and empirical work from some dissertations indicates that other instruments yield lower average scores for older teachers; a belief we have confirmed in our own research where we examined interview scores from in-service teachers from 22-60 years of age. Past 50 years of age, the scores were lower as age increased but we did not find that younger individuals were any different from their middle aged peers. Importantly, some commercial employment screening devices recognize potential differences in age/experience levels and allow the interviewer to select questions for novice or experienced teachers.

Race does seem to matter. When video clips of teachers from different racial groups were viewed by experienced principals, white candidates scored a little better. (5%-10%). We do not fully understand why at this point and do not know if minority principals rate minority candidates differently; however, some adjustment in the scores of minority candidates does seem warranted. Results from similar studies in the private employment sector seem to affirm this observation.

Interviewer gender also seems to be important, but in an unexpected way. Female interviewers tend to rate all candidates regardless of gender slightly more accurately (lower standard deviation) than males. We speculate females have better listening skills but have not investigated this observation fully at this point. Experience or training in a particular field, however, does not seem to improve interviewer accuracy. Experienced principals were no better than experienced teachers in the accuracy of candidate rating.

Importance of Training

Over the last few years we have constructed hundreds of video clips of actors portraying teaching candidates answering interview questions in ways depicting various levels of quality. We have had hundreds of education professionals view these video clips and judge the quality of the teacher actors’ responses to employment interview questions. We have monitored how long it took these educators to make a decision, how many errors they made before reaching the correct answer, and if the actors’ gender, age, race, or attractiveness had any bearing in the decision. Part of our interest in these projects was to help develop questions and check their clarity and potential bias but we also used the video clips to conduct research in a number of areas related to the need for administrator training to conduct employment interviews.

One of our early conclusions was that in general, given quality questions and rubrics, typical teachers and administrators can accurately judge the correct level of response to these actor generated interview clips about 70% of the time with minimal training. Neither administrative experience nor specialized training in a particular field seemed to affect this basic accuracy level. Teacher were just as accurate as veteran administrators when judging the actors playing classroom teachers, and principals were just as accurate as directors of special education when viewing the actors playing potential special education teachers. Practical implications of these findings suggest that the ability to carefully listen to the candidate’s response and compare it to the various rubrics is more important than educational background and experience. Other authors have made similar observations.

A portion of these educators have also completed a three-hour training program designed to familiarize them with the questions, rubrics, and provide extensive practice comparing their responses with known standard video clips. We have used both video tapes and interactive DVD programs in our attempts to obtain mastery level work from these educators. With few exceptions, most educators can achieve 90% accuracy within this 3 hour training period. Indeed, some school

---

districts that currently use some of the instruments we developed mandate that all administrators obtain an accuracy rate of 90% before they are allowed to conduct employment interviews.

Training not only increases individual accuracy rates but also results in higher inter-rater reliability across the given school district. To maintain this high rate of reliability, frequent training is necessary otherwise drift occurs over time. This means that in addition to sound instrument metric qualities as mentioned above, school districts should carefully evaluate the quality and cost of the instrument’s associated training component.

**Predictive Validity of Employment Selection Instruments**

Two questions always seem to surface when discussing the predictive validity of employment screening instruments: are they accurate and is there anything better? The answer to the first question is a yes and no. Yes, given a well-developed set of questions, they will do a credible job separating the excellent from poor teachers. Our studies indicate that within the first 10 questions an 80% accuracy rate can be obtained in the classification of a teacher as falling into the excellent or poor category as identified by his or her principal. Job embedded, situational, and personality type questions all function well in separating the excellent from the poor groups.

Predicting whether a candidate will be an excellent or average teacher is less certain. The error rate increases substantially if you ask a selection instrument to help you make fine level distinctions among candidates. Our research indicates that even the best instrument can only reduce your uncertainty about 25% for cases in the middle. Even with the inclusion of additional information from transcripts, references, background checks, etc. your error margin will still be relatively high. This wide variance is caused by random error, candidate inconsistency in answering the questions, interviewer scoring errors, plus of host of other largely unknown factors. In addition, events occurring after the interview data has been collected can also influence individual teacher development. Experiences with new teacher induction programs, staff development, building acculturation, and interactions with fellow teachers and the principal also play an important role in teacher development.

Employee selection instruments fall somewhere in the middle of the distribution of all data collection strategies for predicting employee future effectiveness. The best single predictor of future job performance is past job performance. Thus, on the job observation, simulations, and apprentice programs are the best way to predict how an employee will do in a new, but similar, position. Having a prospective teacher demonstrate his or her skills in a classroom situation with children through substitute teaching or guest lecturing will provide vastly better information about the skills and abilities of the candidate than any employment selection instrument. Structured interviews generally correlate 0.30-0.60 with subsequent supervisor ratings while on-the-job evaluations are in the 0.70-0.80 range. At the bottom end of the prediction curve are recommendation letters and unstructured interviews that typically correlate less than 0.20 with outcome measures.

**Security Measures**

If you intend to use the employment screening instrument over multiple years, then security becomes an issue. As professors in a large teacher training institute, we have observed that some students already possess interview questions from a given district before they interview with that district. Obviously, this diminishes the predictive validity of the employment interview instrument. When interviews are conducted on-site, question security is less of an issue since the candidate never possesses the set of questions. About the only way they can distribute the questions to others is by memorizing them, which is difficult.

In contrast, if you purchase or develop an employment interview distributed over the Internet, then security is problematic for two reasons. First, anything appearing on a computer screen can be captured to the hard drive, printed, and e-mailed to anyone who might have an interest. The process is relatively simple (press “print screen” and then save to a word processing program). Thus, candidates can take the test for one district, save the questions, figure out the best answers, and then re-take the identical test from another district. The other issue is that one can never be sure who is completing the test. It might be the candidate, the candidate’s friends, or a group of individuals.
Question security can be enhanced in a number of ways. The easiest way is develop a bank of questions to use and then randomly select different questions for each interview. If the same questions are routinely used, then the interview instrument should at least present the questions in a random order each time. Distribution over the Internet probably requires random selection and presentation of questions from a large bank with time limitations for each question. An alternative is to design computer adaptive testing process similar to the one used by ETS on the SAT. Before you develop or purchase a commercial employment interview instrument, carefully consider these security issues.

**The Bottom Line**

Three questions seem relevant. First, is the cost and training time involved with the purchase of a structured employee interview system worthwhile? Second, should your district develop its own employment selection screening instrument? Third, if you decide to purchase one of the commercial instruments, which one is best?

The answer to the first question is an unequivocal “yes”. Employing well developed and psychometrically sound structured interviews in the selection process will indeed double your chances of identifying excellent teachers. There has been over thirty years of research both in education and business to support this conclusion. The cost benefit analysis of on-line questionnaires, personality profiling systems, and interest inventories is less certain since these internet versions have not been as extensively examined.\(^\text{10}\)

The answer to the second and third questions is “it depends”. If your district has the expertise and time to develop the procedures, questions, rubrics, and training materials including video taped vignettes, then by all means consider building an instrument yourself. If you follow the recommendations in the NCME guidelines, you likely will construct an instrument that is at least equivalent to the commercially developed ones. There is no magical quality to the commercial products. If, your district lacks technical expertise or time, then selecting from the available commercial instruments is probably best.

All of the commercially produced instruments will assist school districts in selecting quality teachers primarily because they are all structured instruments, which far exceed non-structured or question only instruments common in many school districts. Primarily, what you need to decide is if the particular instrument measures the attributes you want in your incoming employees, whether the instrument is reliable and valid, if adequate training is available, and obviously the cost. Do remember, however, that even the best commercial instrument can only predict teacher quality about 25% of the time so be cautious in your interpretation of those seemingly scientific scores.\(^\text{11}\)

---

\(^{10}\) Published research on the Gallup on-line instrument is meager and inconclusive. Some authors have discovered respectable correlations with principal and student ratings while others studies conclude there is little correlation. We could find no available research on the Ventures or Haberman on-line instruments.

\(^{11}\) The percent of variance predicted by the instrument is the square of the correlation coefficient between the selection instrument and measured teacher quality. Even the best of the commercial instruments do not claim that relationship is very high.