



The Research Behind the ETS[®] Personal Potential Index (PPI)

Patrick C. Kyllonen
Educational Testing Service

A Background Paper From ETS

Introduction

The ETS® Personal Potential Index (PPI) is a rating system for assessing a graduate school applicant's suitability for graduate study. The purpose of this white paper is to review the basis for the PPI. The paper reviews the history of GRE® involvement with noncognitive factors, various methods that have been used to measure noncognitive factors, and evidence for which factors are most important to measure. The paper also reviews the predecessor to the PPI, the Standardized Letter of Recommendation, which ETS used operationally to select graduate student interns, and which is currently being used in Project 1000. Finally, the paper suggests future directions for how PPI scores could be adjusted for increasing validity.

GRE History with Noncognitive Assessments

There has been a long history of interest in noncognitive factors by ETS (Kyllonen, 2005a), and more specifically, the GRE Board (e.g., Enright & Gitomer, 1989). But the more immediate stimulus for the current effort is GRE's Horizons project. Horizons was undertaken in the late 1990s to understand the assessment needs of graduate institutions using the GRE, to evaluate how well the GRE tests met those needs, and to recommend changes (Briel, Bejar, Chandler, Powell, Manning, et al., 2000). Horizons initially involved interviews with graduate school deans and faculty at 61 institutions. A follow-up to that project designed to define successful graduate students involved more in-depth telephone interviews with 16 faculty members and five deans from five institutions (Walpole, Burton, Kanyi, & Jackenthal, 2001).

An analysis of the transcripts produced a noteworthy finding. Certain noncognitive factors, such as persistence, collegiality, and communication, were among the most frequently mentioned important qualities, even more frequently mentioned than mastery of the discipline and ability to teach. Horizon interviewees also expressed concern about the continued problem of graduate school attrition and time-to-degree, which they suggested may at least partly be accounted for by noncognitive factors.

The Horizons findings led the GRE Board to organize a special session at its 2001 Toronto meeting, entitled "Symposium on Noncognitive Traits and Assessments in the Context of Graduate Admissions." Several prominent researchers — Lew Goldberg, Sandra Graham, Milton Hakel, Charlie Reeve, Jack Mayer, and William Sedlacek — presented ideas and findings to the board on what noncognitive factors might be most important to graduate school success and how they might best be measured. Goldberg reviewed the Big 5 personality framework, Graham discussed self-efficacy and attribution theory, Reeve and Hakel discussed contributions from industrial-organizational psychology, including the research on faking in self-assessments, Mayer discussed emotional intelligence, and Sedlacek reviewed a variety of experimental methods that had been employed to assess noncognitive factors. The ideas presented elicited animated discussion, but in the end the board decided that noncognitive assessment for the GRE was premature. Assessments for which there was validity evidence (e.g., self-assessments) seemed susceptible to faking, and assessments that might be resistant to faking (e.g., emotional intelligence measures) lacked validity evidence. The Board commissioned a literature review, and recommended that the issue of noncognitive assessment be revisited if significant new developments occurred or new findings came to light.

Methods for Measuring Noncognitive Factors

There are a variety of methods for measuring noncognitive factors. These include self-assessments, situational judgment tests, biographical reports, objective measures, such as reaction time tests, and others' ratings.

Self-assessments are most common, and are responsible for most of what we know about the relationship between noncognitive factors and educational and employment outcomes. Several meta-analyses have been conducted on such relationships, based almost exclusively on self-assessments (e.g., Barrick, Mount, & Judge, 2001). The problem with self-assessments is that respondents can fairly easily fake their responses to appear more attractive to a prospective employer or institution. ETS has conducted several mini-conferences addressing the faking problem,¹ and has identified several promising methods for collecting self-assessments, such as the multidimensional forced choice format (pitting equally attractive noncognitive factors — such as “works hard” and “works well with others” against each other), but evidence for their utility is still mixed (Heggestad, Morrison, Reeve, & McCloy, 2006).

Situational judgment tests (SJTs) present a scenario posing some kind of problem, and the examinee is asked how best to solve the problem. Typical item formats are either multiple choice or Likert scale ratings (often five-point) of multiple presented solutions. The SJT format permits both text and video-based tests, it has been researched widely (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; Sternberg, Forysthe, Hedlund, Horvath, Wagner, et al., 2000) and is commonly used in industry. SJTs have been developed to measure a wide variety of constructs, including leadership, the ability to work with others, achievement orientation, self-reliance, dependability, sociability, agreeableness, social perceptiveness, and conscientiousness (e.g., Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004; Waugh & Russell, 2003). Hence, SJTs can measure noncognitive as well as cognitive qualities (McDaniel & Nguyen, 2001), particularly if examinees are asked to indicate how they would respond (which tends to reflect personality) rather than what might be the best response (which tends to reflect ability). ETS has developed numerous SJTs, ranging from print-based measures of business analysis and problem solving (Kyllonen & Lee, 2005), to video-based measures of communication skills (Kyllonen, 2005b). In the noncognitive realm, ETS is currently developing SJTs in the areas of leadership and social competence. Although SJTs may be less susceptible to faking, research has shown that applicants seem to be able to improve their score if allowed to retest (Lievens, Buyse, & Sackett, 2005), or after being coached (Cullen, Sackett, & Lievens, in press), but only by about a quarter of a standard deviation. This is much lower than comparable effects for self-assessments. SJTs are so variable in style and format, as well as in the construct they measure, that it could well be the case that some not-yet-determined design specifics govern their susceptibility to faking.

¹ These include two mini-conferences on alternative methods to measure personality (Roberts, Schulze, & Kyllonen, 2006a [see Holden, 2006; Paulhus, 2006; Prelec & Weaver, 2006; Sackett, 2006; Zickar, 2006]; Roberts, Schulze, & Kyllonen, 2006b [see Ellingson, 2006, Heggestad, 2006, Lukoff, 2006, Hancock, 2006, Zeigler, 2006]), and two conference symposia (Kyllonen, 2006 [see Cvencek & Greenwald, 2006; Kubinger, 2006; Prelec, 2006; White, Young, Stark, Drasgow, & Hunter, 2006]; Kyllonen, 2007 [see Kuncel, 2007, Lukoff, 2007, Seiler, 2007, Stark & Chernyshenko, 2007, Ziegler, 2007]).

Verifiable biodata measures are self-reports of noncognitive factors that are, in principle, observable and they may be accompanied by requests to provide verification. For example, respondents may be asked, “How many times did you lead class discussions during your senior year in high school?”, followed by a request to list the classes and topics; or, “In how many different languages besides English can you converse well enough to order a meal?”, followed by a request to list the languages (Schmitt, Oswald, Gilespe, Ramsay, & Yoo, 2003). Stricker, Rock, & Bennett (2001) developed a documented accomplishments measure that produced scores for six scales: Academic Achievement, Leadership, Practical Language, Aesthetic Expression, Science, and Mechanical. For the Leadership category, items were: *Was on a student-faculty committee in college. Yes/No. If YES: Position, organization, and school?* This approach is currently under investigation by the College Board in a project on undergraduate admissions.

Objective measures. There are several experimental measures being investigated that fall into the category of performance or objective noncognitive measures. *Conditional reasoning tests (CRTs)* measure reasoning conditioned on *worldview*, or unstated assumptions and preconceptions (James, 1998). It is assessed through a set of five-alternative multiple-choice reading comprehension-type items where three of the alternatives are logically wrong, and the other two logically correct. The two correct alternatives vary in the worldview underlying them, and so an examinee’s selection of one of the two can be treated as an indicator of the worldview of that examinee. James (1998) has developed conditional reasoning tests for achievement motivation and aggression, and some versions are being marketed by Harcourt Assessment, Inc. However, the method has proven difficult to replicate (Gustafson, 2004). *Implicit association tests (IATs)* are a very different kind of objective test that measure the naturalness of an association between two concepts for an individual by comparing reaction times (Greenwald, McGhee, & Schwartz, 1998). These have been enormously popular in social psychology research (Nosek, Greenwald, & Banaji, 2005), and have been used to measure noncognitive factors, but thus far have not been evaluated in a selection or admissions context.

Others’ ratings are assessments in which raters (e.g., supervisors, trainers, colleagues, friends, faculty advisors) rate applicants on various noncognitive qualities. There is a long history for using this method (e.g., Tupes and Christal, 1961/1992). Because socially desirable responding is taken out of the hands of the applicant, this method is often assumed to be free of faking, and can even be (and often is) used as a criterion against which to evaluate the effects of faking on self-assessments. However, it can also be used as a measure of applicant personality in its own right. For example, ETS developed and for several years has administered a predecessor to the PPI called the Standardized Letter of Recommendation (SLR), which was completed by faculty raters to measure applicants’ noncognitive qualities for selection into summer intern and fellowship programs (Kyllonen & Kim, 2004; Walters, Kyllonen, & Plante, 2003; 2006). Based on the fact that others’ ratings are the most likely of the methods reviewed to minimize faking, combined with the fact that they have been used operationally at ETS and through Project 1000 to select graduate student applicants, we believe that the others’ ratings method is the only viable approach to measuring noncognitive factors in a high-stakes context at this time.

Most Important Noncognitive Factors to Measure

There are several data sources for identifying which noncognitive qualities are most important to measure. One of these is interviews of graduate faculty and staff members. The Enright and Gitomer (1989) study based on over-the-phone interviews and follow-up group discussions yielded seven general competencies. Several of these were noncognitive — communication (ability to share ideas, knowledge, and insights with others), motivation (commitment, involvement, interest in work), planning and self-organization (developing a procedure to reach a goal), and professionalism (skills in accommodating to social conditions in field).

The Horizons follow-up study (Walpole, et al., 2001), also based on interviews, asked faculty to identify variables that ought to be considered in admissions as predictors. A number of noncognitive variables rated quite highly, including values and character, maturity and responsibility, and interpersonal skills. All these were mentioned even more often than research experience and problem-solving ability. In addition to these there were other noncognitive variables mentioned, including (in order of mentions) persistence, collegiality, communication, independence, creativity, enthusiasm, values/character and open-mindedness.

A second source for identifying key noncognitive qualities is the “taxonomy of higher order performance components” based on a correlational analysis of training and performance outcome measures conducted by the U.S. Army (“Project A,” Campbell, 1990; Campbell, et al., 1993). The eight taxonomic factors are intended to be distinct and comprehensive and to represent of the range of outcomes that organizations value in their employees. Two of these — job-specific proficiency, non-job-specific proficiency — are cognitive. But the remaining six are primarily noncognitive — written and oral communication, demonstration of effort, maintenance of personal discipline, facilitation of peer and team performance, supervision and leadership and management and administration. This system has been widely adopted in industrial-organizational psychology (e.g., see Bartram’s [2005] “Great Eight”). Also, Kuncel, Hezlett, & Ones (2001) presented a variant on this model to accommodate undergraduate student performance, based on critical incidents. Reeve and Hakel (2001) presented an “armchair” adaptation of the Kuncel, et al. model to accommodate graduate student performance.

A recent widely cited survey that incorporated these factors was administered to U.S. employers to determine workforce readiness (The Conference Board, et al., 2006). It included the factors of critical thinking and problem solving, teamwork and collaboration, oral and written communication, work ethic, leadership, information technology, diversity, lifelong learning, creativity and ethics.

A question is whether these lists are somewhat arbitrary and endless. However, personality psychology (e.g., Goldberg, 1990) suggests that individual differences in noncognitive factors are organized along only five independent dimensions — *extroversion*, *agreeableness*, *emotional stability*, *conscientiousness*, and *openness* (there is a new, closely related theory that posits a sixth factor, *honesty*; Saucier, in press). Of these, *conscientiousness* and *emotional stability* are most commonly found to be predictive of performance outcomes (Barrick, et al., 2001).

Item Development for the PPI

Based on our review (Kyllonen et al., 2005; 2006) suggesting that of the various methods, only others' ratings were free of the problems of faking, we developed an others' rating system initially called the SLR. The system was designed to combine aspects of a traditional letter of recommendation with those of a rating form. Like a traditional letter, the system enabled comments (along with the ratings), and asked faculty members to compare the student with others they had supervised. But the system was more structured than a conventional letter, prompting the faculty member to respond to specific items using a Likert rating scale (details below). The design of the system was somewhat iterative, beginning with a series of focus groups, including discussions with GRE Board members, followed by some changes in its form, followed up with a large-scale ($N = 400$) telephone survey of faculty members (Walters, Kyllonen, & Plante, 2003; 2006). An analysis of the large-scale study results was conducted using both classical and Rasch measurement approaches (Kyllonen & Kim, 2006; Kim & Kyllonen, 2006). The form of the survey resulting from this initial effort (and several minor variants) was adopted for operational use in ETS's summer intern and fellowship award programs (Liu, Minsky, Ling, & Kyllonen, 2007). Following this, the survey and reporting procedure was modified for use in Project 1000. Modifications were made both to the content and the functionality based on Project 1000's requirements.

Factors and Items

Seven scales comprised the initial SLR — *Knowledge & Skill*, *Creativity*, *Communication Skills*, *Teamwork*, *Motivation*, *Self-Organization* and *Professionalism & Maturity*. These scales were identified based primarily on the findings from the faculty surveys (Enright & Gitomer, 1989; Walpole, et al., 2001), along with considerations from the larger literature review (Kyllonen, et al., 2005; 2006), and the faculty interviews and focus groups (Walters, Kyllonen, & Plante, 2003; 2006). We decided to measure each of these scales with four items, yielding 28 items. Items themselves were written with the assistance of two expert personality psychologists, Lewis Goldberg and Gerard Saucier. Scales and items were reviewed and revised through focus groups and usability studies (Walters, et al., 2003; 2006).

Rating Scale

Several rating scale formats were initially considered, including a behaviorally anchored rating scale. However, for several of the scales we found it difficult to develop behavioral anchors specific enough to serve as good examples, but general enough to span the variety of graduate disciplines (e.g., masters vs. doctoral degree; humanities vs. sciences). So instead we opted for a Likert rating scale of items.

Raters are asked to:

“Please rate the applicant relative to others in your department or unit who have gone on to study at one of the participating institutions.”

Ratings were on a 1–5 scale, with 1 = below average; 2 = average, 3 = above average; 4 = outstanding, 5 = truly exceptional. There was also a “do not know” option. The scale labels were imbalanced in allowing for three levels above average, but only one below. This was designed to reflect ratings inflation. Empirical analyses to date suggest that this strategy was successful. We have consistently found a mean of between 3 (above average) and 4 (outstanding), with a standard deviation of approximately 1 for the various items.

Empirical Findings & Revisions

There have been two sets of empirical analyses of the SLR. The first (Kyllonen & Kim, 2005; Kim & Kyllonen, 2006) was based on the analyses of ratings from the 430 faculty members who were asked to rate the last person for whom they already completed a letter of recommendation. The key findings were that faculty members differentiated students, that there was a strong halo, but there was some evidence for multidimensionality. Differentiation was shown both by the lack of a ceiling effect (item means were all less than 4, with standard deviations slightly less than 1), and by the high reliabilities both overall ($\alpha = .97$) and by scale (alphas ranged from .84 to .89 for the four-item scales). There was a strong overall rating factor, or halo, such that some students were rated more highly across the board than were others. This was shown by the high overall reliability, and by the strength of the first factor, which accounted for 80 percent of the common variance among items. However, there was also evidence that faculty members were not only rating students along one general dimension, but were making distinctions between different qualities. Factor analytic results were consistent with the idea that faculty members differentiated between cognitive and noncognitive qualities, and that within the noncognitive realm, they reliably differentiated Teamwork, Professionalism, Motivation, and Communication Skills. Knowledge and Creativity were not differentiated, which led to our changing the format to combine knowledge and creativity items into one scale.

The second set of empirical analyses was based on actual operational data provided by professors' ratings of students applying for an ETS internship program. Data were accumulated over three years and were based on 414 ratings of intern applicants. We also asked ETS research mentors to complete the rating form (the SLR) on those summer interns who were selected and worked under the supervision of that mentor. Only a small proportion of applicants are typically selected, and so only 51 ratings were completed by mentors over the three-year span. There were two major findings. First, although ratings from the operational study were higher than those from the non-operational study, the difference was only about a half a point. It is not clear whether this reflects differences in the qualities of the applicants or differences in the amount of inflation due to it being an operational assessment. But the somewhat surprising and welcome finding was that even under the operational press of internship selection, the ratings still did suffer exceedingly from a ceiling effect (see Figure 1; or compare Tables 1 and 2).

Although the small N of mentor ratings prevents making too much out of the validity data, it is noteworthy that the correlation between faculty ratings and ETS mentor ratings was fairly high, ranging from $r = .41 - .60$ except for two scales with correlations in the .20s, which also happened to have the highest means, and may therefore have suffered from range restriction due to a ceiling effect. These correlations are particularly high given previous ratings research. For example, Baxter, Brock, et al., (1981) reported fairly low inter-rater reliability ($r = .40$). And a meta-analysis by Vanelli, Kuncel, & Ones (2007) found correlations between letters of recommendation (holistically scored) and faculty ratings of only $r = .25$ (other correlations with letters of recommendation were $r = .10$ [with research productivity], $.14$ [graduate GPA], $.19$ [Ph.D. attainment], $.28$ [undergraduate GPA], and $.25$ [faculty ratings]).

Factor analytic results from the two sets of data, along with the findings of a near ceiling effect for some of the items in the operational study, suggested that some items were not effective as they could be. In consultation with an expert personality psychologist, Gerard Saucier, we also noted that the rating scale lacked items pertaining to an applicant's emotional responsiveness. We therefore made adjustments to the scale by deleting several items, modifying the wording of several, and adding two items to measure emotional responsiveness: "accepts feedback without getting defensive," and "works well under stress." The final set of items is shown in Table 3.

This final set of items (Table 3) was incorporated into a revised version of the SLR, accompanied by a name change to the PPI. The PPI is currently being pilot tested through Project 1000, a national program designed to assist underrepresented students applying to graduate school. Project 1000 is a centralized application processing service. Students apply to up to seven graduate schools through Project 1000. Project 1000 collects PPI ratings from student-nominated faculty referees, prints out reports summarizing those ratings, bundles the reports with transcripts and other application materials, then mails out packages to those graduate programs to which the students are applying. We will conduct analyses on data back from the Project 1000 pilot study during the spring and summer of 2008. We expect to demonstrate that the items and scales are sufficiently reliable to support continued use, and we expect to be able to provide psychometric evidence for the validity of the system, for example through factor analysis, and perhaps some outcome correlations (e.g., comparing selected with unselected students). Over time we would like to evaluate predictive validity against other criteria, such as retention, time to degree, graduate grade-point average, faculty ratings, and so on. We also will survey users (referees and admissions committee members) to evaluate PPI usability and users' satisfaction with the PPI.

Scoring and Reporting Adjustments for the PPI

The design and format of the PPI rating system presents potential interpretation problems due to biases caused by raters not rating all applicants. Applicants with lenient raters will get higher scores due to the effect of the rater, independent of the qualities of the applicant. The recipient of the ratings could be unaware of this kind of effect, for ratings in general, or for an applicant (or rater) in particular.

One could argue that this is no different or no worse than the problem currently presented by conventional letters of recommendation, or even by grades. So in the first version of the PPI there was no attempt to adjust ratings to take any such biases into account.

However, there are several adjustments that could be made and evaluated in future research studies:²

National and Local Norm Anchoring

A problem with noncognitive factors is that users do not have the same kind of history with them that they have with other factors such as grades, or even the evaluations presented in conventional letters of recommendation. As experienced recipients of grades and letters, senior faculty members are capable of informal adjustments of incoming data based on past history. An A average from School X may be seen as no better than a B average from school Y, and a moderate letter from Professor N may be seen as equivalent to a strong letter from Professor M. While some of this informal adjustment by recipient could happen with the PPI, additional support could also be provided, in principle, and over time. One form of additional support would be to provide national and local norms, either (or both) to the referee or to the score recipient. National or local percentiles could be provided for each item (or scale). For example, for the item "Works well in group settings," the national 25th, 50th, and 75th percentiles could be provided (e.g., 3.3, 3.9, 4.4), or the university's chemistry department percentiles could be provided (for the department of the rater). Local norms could also be passed forward (or not) to the score recipient. Or the ratings could be scaled on a common scale prior to transmission to the receiving institution (then the receiving institution could readjust ratings to take into account the strength of applicants from that university's department). Obviously, this is not achievable right away, but the value of such data would increase over time with more data entered into the system.

²These suggestions have been developed and discussed with Dan Koretz, David Payne, and other GRE Board members and ETS staff members.

Instant Rater History Feedback

This is a fairly simple idea involving only a private record of a referee's ratings history. Once a referee completes a rating (of an item, or a scale, or overall), the system would compare the referee's past ratings, which would be locally and privately stored, and would provide feedback, such as "your rating just placed this applicant at your 90th percentile, is that what you wanted to do?" The purpose of this feedback would be to provide a kind of reality check to the rater. The system might also allow the referee to review his or her ratings history, to check for appropriate spread and other patterns.

Rater Severity Adjustments

In rating studies, severity adjustments are ones to rescale one rater's ratings to be more in line with (on the same scale as) ratings from other judges — translating a severe rater's ratings into the rating scale used by the others. These kinds of adjustments are easy to make when a set of raters rates a common set of applicants. With letters of recommendation (or with the PPI), this is not entirely the case. There is not a lot of overlap between different referees and different applicants, particularly across schools, but there is some. A design may be able to be worked out to take advantage of the overlap that exists to enable severity adjustments.

Rater Validity Adjustments

Rater validity adjustments are like rater severity adjustments, except that a rater's ratings are scaled to align with applicant outcomes, such as completing school, or graduate GPA, or any other outcome that could be collected. Thus, such rater-validity-adjusted ratings take into account the predictive validity of a referee's ratings in predicting outcomes, and puts the ratings from different raters on a common scale aligned to those outcomes. It is not clear how practical such an adjustment would be, at least initially, but over time, and if outcomes data of some kind were routinely collected, such an adjustment would be possible. This kind of adjustment (along with rater severity adjustments, discussed above) would be used by score recipients.

Other Types of Validity

Predicting outcomes is an important form of validity evidence, but there are other sources of validity evidence. For example, validity evidence supporting the content of the PPI could be established by asking faculty members to evaluate PPI scales and items. We will conduct several studies in the upcoming year to establish such evidence. We will convene an expert panel comprised of graduate deans, who will gather to discuss whether the particular qualities measured in the current version of the PPI — *Knowledge and Creativity, Communication Skills, Teamwork, Resilience, Planning and Organization* and *Ethics and Integrity* — are the most important qualities to be measuring for graduate school, and if not, we will discuss gaps, omissions, and wording changes. We also will discuss the items used on the PPI (see Table 3) to determine if these capture the qualities deemed most important. Once this exercise is completed, we will conduct a follow-up study with faculty members in the form of an online survey, addressing these and similar questions.

Another form of validity is consequential validity — for example, teaching, learning, and graduate-school cultural effects resulting from the use of the PPI. For example, PPI qualities such as *Teamwork*, if used as part of an admissions system, might over time be expected to be perceived as more important, resulting in more attention given to them in graduate school preparation activities. Evaluation of these kinds of effects is a more long-term objective.

Summary

Extensive surveys and other research findings strongly suggest that noncognitive factors are important for success in graduate education. However, because of the problem of faking with self-assessments, and the unproven validity of other kinds of noncognitive assessments, these approaches are not viable. The only viable noncognitive assessment system available now is an others' rating system, such as the PPI. The PPI has the additional benefit of having been proven in operational selection for graduate students, first as part of ETS's summer intern program, and currently as a component of Project 1000.

Use of the PPI presents many opportunities for improving student performance, and also presents many research challenges. Some of the key questions are: What is the criterion-related validity of the PPI, using grades, retention, and other standard GRE criteria? What is the construct validity of the PPI, based on factor analyses? Is there a justification for subscores, or should there be one "noncognitive index?" Additional questions are: How can rater effects be controlled; that is, what is the meaning of ratings? How can the consequential validity of the PPI be examined? Do users find rater information from the PPI useful? Does the PPI cover the most important student qualities? How do we know? These are the kinds of questions that can be addressed over the first few years of use with the PPI.

Table 1. Descriptive Statistics for the Professor Ratings of the SLR

Scale/Item	N	Mean	SD	Scale Mean	Scale SD
Knowledge					
1. Has foundational knowledge in some branch of behavioral science (e.g., educational theory, psychology, economics).	403	3.70	.68	3.63	.59
2. Has knowledge of behavioral science research methods.	406	3.79	.66		
3. Has knowledge of educational or psychological assessment.	394	3.77	.74		
4. Has knowledge of program evaluation.	354	3.25	.90		
Analytical Skills					
5. Has quantitative research methods and data analysis skills.	410	3.97	.75	3.75	.63
6. Has qualitative research methods and data analysis skills.	360	3.27	1.00		
7. Has demonstrated skill in measurement and assessment techniques.	388	3.80	.74		
8. Is skilled in statistical methods.	402	3.88	.79		
Communication Skills					
9. Demonstrates clear and critical thinking.	414	4.05	.63	3.83	.64
10. Speaks in a clear, organized, and logical manner.	414	3.74	.80		
11. Writes with precision and style.	410	3.51	.79		
12. Listens well and responds appropriately.	414	4.02	.69		
Motivation					
13. Maintains high standards of performance.	414	4.22	.61	4.25	.52
14. Can overcome challenges and setbacks.	408	4.15	.61		
15. Seek out opportunities to learn.	414	4.38	.59		
16. Has a high level of energy.	413	4.25	.60		
Self-Organization					
17. Organizes work and time effectively.	411	4.06	.64	4.02	.54
18. Sets realistic goals.	408	3.85	.63		
19. Makes good decisions.	407	3.95	.62		
20. Can work independently of others.	412	4.21	.60		
Professionalism & Maturity					
21. Maintains high ethical standards.	408	4.30	.58	4.28	.52
22. Demonstrates honesty and sincerity.	412	4.34	.57		
23. Is dependable.	414	4.35	.59		
24. Regulates own emotions appropriately.	407	4.17	.61		
Teamwork					
25. Shares ideas easily.	412	4.02	.63	4.15	.55
26. Supports the efforts of others.	407	4.12	.63		
27. Works well in group settings.	408	4.15	.65		
28. Behaves in an open and friendly manner.	414	4.32	.61		

Table 2. Descriptive Statistics for the ETS Mentor Ratings of the SLR

Scale/Item	N	Mean	SD	Scale Mean	Scale SD
Knowledge					
1. Has foundational knowledge in some branch of behavioral science (e.g., educational theory, psychology, economics).	51	3.31	.73	3.15	0.70
2. Has knowledge of behavioral science research methods.	51	3.33	.87		
3. Has knowledge of educational or psychological assessment.	51	3.11	.91		
4. Has knowledge of program evaluation.	51	2.70	.98		
Analytical Skills					
5. Has quantitative research methods and data analysis skills.	51	3.49	.84	3.30	0.78
6. Has qualitative research methods and data analysis skills.	51	3.42	1.03		
7. Has demonstrated skill in measurement and assessment techniques.	51	3.12	.97		
8. Is skilled in statistical methods.	51	3.25	1.01		
Communication Skills					
9. Demonstrates clear and critical thinking.	51	3.44	1.01	3.39	0.91
10. Speaks in a clear, organized, and logical manner.	51	3.48	.91		
11. Writes with precision and style.	51	3.19	.85		
12. Listens well and responds appropriately.	51	3.50	1.09		
Motivation					
13. Maintains high standards of performance.	51	3.58	.86	3.55	0.92
14. Can overcome challenges and setbacks.	51	3.48	1.03		
15. Seek out opportunities to learn.	51	3.58	.98		
16. Has a high level of energy.	51	3.58	1.03		
Self-Organization					
17. Organizes work and time effectively.	51	3.34	1.04	3.35	0.89
18. Sets realistic goals.	51	3.29	.92		
19. Makes good decisions.	51	3.30	.84		
20. Can work independently of others.	51	3.52	1.02		
Professionalism & Maturity					
21. Maintains high ethical standards.	51	3.57	.89	3.54	0.89
22. Demonstrates honesty and sincerity.	51	3.58	.90		
23. Is dependable.	51	3.50	1.03		
24. Regulates own emotions appropriately.	51	3.60	.91		
Teamwork					
25. Shares ideas easily.	51	3.33	1.01	3.43	0.85
26. Supports the efforts of others.	51	3.35	.94		
27. Works well in group settings.	51	3.45	.86		
28. Behaves in an open and friendly manner.	51	3.67	.92		

Figure 1. Professor Rating

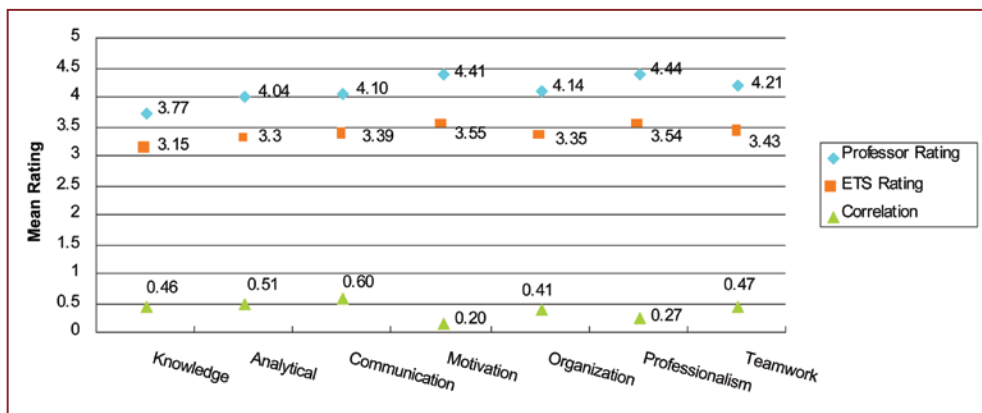


Table 3. PPI Scales and Items

Knowledge and Creativity

Has a broad perspective on the field
 Is among the brightest persons I know
 Produces novel ideas
 Is intensely curious about the field

Communication Skills

Speaks in a clear, organized, and logical manner
 Writes with precision and style
 Speaks in a way that is interesting
 Organizes writing well

Teamwork

Supports the efforts of others
 Behaves in an open and friendly manner
 Works well in group settings
 Gives criticism/feedback to others in a helpful way

Resilience

Accepts feedback without getting defensive
 Works well under stress
 Can overcome challenges and setbacks
 Works extremely hard

Planning and Organization

Sets realistic goals
 Organizes work and time effectively
 Meets deadlines
 Makes plans and sticks to them

Ethics and Integrity

Is among the most honest persons I know
 Maintains high ethical standards
 Is worthy of trust from others
 Demonstrates sincerity

References

- Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). The FFM personality dimensions and Job performance: Meta-Analysis of Meta-Analyses. *International Journal of Selection and Assessment*, 9, 9 – 30.
- Bartram, D. (2005). The Great Eight Competencies: A criterion-centric approach to validation. *Journal of Applied Psychology*, 90, 1185 – 1203.
- Baxter, J. C., B. Brock, et al. (1981). Letters of recommendation: A question of value. *Journal of Applied Psychology*, 66(3), 296 – 301.
- Briel, J., Bejar, I., Chandler, M., Powell, G., Manning, K., Robinson, D., Smallwood, T., Vitella, S., & Welsh, C. (2000) *GRE Horizons Planning Initiative*. (Graduate Record Examination). A research project funded by the GRE Board Research Committee, the GRE Program, and the Educational Testing Service Research Division.
- Campbell, J. P. (1990). An overview of the army selection and classification project (Project A). *Personnel Psychology*, 43, 231 – 239.
- Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. Borman (Eds.), *Personnel selection in organizations*. San Francisco: Jossey-Bass.
- Cullen, M. J., Sackett, P. R., & Lievens, F. (In press). Threats to the operational use of Situational Judgment Tests in the college admission process. *International Journal of Selection and Assessment*.
- Cvensek, D. & Greenwald, A. (2006). Implicit Association Test. In P. C. Kyllonen (organizer), Solving the faking problem on noncognitive assessments (Invited symposium: Psychological Assessment and Evaluation). Athens, Greece: *26th International Congress of Applied Psychology*.
- Ellingson, J. (2006). Personality retest effects: Guilt as a mechanism for managing response distortion. In R. D. Roberts, R. Schulze, & P. C. Kyllonen (chairs), *Technical Advisory Committee on Faking on Noncognitive Assessments*. Princeton, NJ: ETS.
- Enright, M. K., & Gitomer, D. (1989). *Toward a description of successful graduate students* (GRE Board Research Report No. 85-17R). Princeton, NJ: ETS.
- Goldberg, L. R. (1990). An alternative “description of personality”: The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59, 1216 – 1229.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464 – 1480.
- Gustafson, S. (Chair) (2004). *Making conditional reasoning test work: Reports from the frontier*. Symposium conducted at the 19th Annual Conference of the Society for Industrial and Organizational Psychology Conference, Chicago, IL.
- Hancock, J. (2006). Situational and psychological factors predicting deception and its detection (in noncognitive assessment). In R. D. Roberts, R. Schulze, & P. C. Kyllonen (chairs), *Technical Advisory Committee on Faking on Noncognitive Assessments*. Princeton, NJ: ETS.
- Heggstad, E. (2006). Faking in personnel selection: Does it matter and can we do anything about it? In R. D. Roberts, R. Schulze, & P. C. Kyllonen (chairs), *Technical Advisory Committee on Faking on Noncognitive Assessments*. Princeton, NJ: ETS.
- Heggstad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology*, 91, 9 – 24.
- Holden, R. R. (2006). *Faking on noncognitive self-report: Seven primary questions*. Paper presented at the ETS Mini-conference on Faking in Noncognitive Assessments. Princeton, NJ: ETS.
- James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods*, 1(2), 131 – 163.
- Kim, S., & Kyllonen, P. C. (2006) *Rasch rating scale modeling of data from the Standardized Letter of Recommendation*. ETS Research Report RR-06-33. Princeton, NJ: ETS.

- Kubinger, K. D. (2006). Is there no way out of job candidates' faking good on personality questionnaires? In P. C. Kyllonen (organizer), *Solving the faking problem on noncognitive assessments* (Invited symposium: Psychological Assessment and Evaluation). Athens, Greece: *26th International Congress of Applied Psychology*.
- Kuncel, N. (2007). The desirability of item response options and their effect on faking behavior. In P. Kyllonen (chair), *The faking problem in noncognitive assessment*. Symposium presented at the American Psychological Association Annual Convention, San Francisco, CA.
- Kyllonen, P. C. (2005a). *The case for noncognitive assessments* (R&D Connection No. 4). Princeton, NJ: ETS. Available at: <http://www.ets.org/portal/site/ets/menuitem.c988ba0e5dd572bada20bc47c3921509/?vgnnextoid=7f8de4a8d1076010VgnVCM10000022f95190RCRD&vgnnextchannel=dcb3be3a864f4010VgnVCM10000022f95190RCRD>
- Kyllonen, P. C. (2005b). Video-based communication skills test for use in medical college. In N. Gafni (Organizer), *Assessment of noncognitive factors in student selection for medical schools*. Symposium conducted at the annual meeting of the American Educational Research Association, Montreal, Canada.
- Kyllonen, P. C. (2006). *Solving the faking problem on noncognitive assessments* (Invited symposium: Psychological Assessment and Evaluation). Athens, Greece: *26th International Congress of Applied Psychology*
- Kyllonen, P. C. (2007). *The faking problem in noncognitive assessment*. Symposium presented at the American Psychological Association Annual Convention, San Francisco, CA.
- Kyllonen, P. C., & Lee, S. (2005). Assessing problem solving in context. In O. Wilhelm & R. W. Engle (Eds.) *Handbook of Understanding and Measuring Intelligence* (pp. 11 – 25). Thousand Oaks, CA: Sage.
- Kyllonen, P. C., Walters, A. M., & Kaufman, J. C. (2005). Noncognitive constructs and their assessment in graduate education: A review. *Educational Assessment*, 10, 153 – 184.
- Kyllonen, P. C., Walters, A., & Kaufman, J. (2006). *Noncognitive constructs in graduate education* (GRE Board Report No. 00-11R). Princeton, NJ: ETS.
- Kyllonen, P. C., & Kim, S. (2005). *Personal Qualities in Higher Education: Dimensionality of Faculty Ratings of Graduate School Applicants*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.
- Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest Effects in operational selection settings: Development and Test of a Framework. *Personnel Psychology*, 58, 981 – 1007.
- Liu, O. L., Minsky, J., Ling, G., & Kyllonen, P. (2007). *The Standardized Letter of Recommendation: Implications for Selection*. ETS Research Report RR-07-38. Princeton, NJ: ETS.
- Lukoff, B. (2006). Using decision trees to detect faking in noncognitive assessments. In R. D. Roberts, R. Schulze, & P. C. Kyllonen (chairs), *Technical Advisory Committee on Faking on Noncognitive Assessments*. Princeton, NJ: ETS.
- Lukoff, B. (2007). Detecting faking on noncognitive assessments using decision trees. In P. Kyllonen (chair), *The faking problem in noncognitive assessment*. Symposium presented at the American Psychological Association Annual Convention, San Francisco, CA.
- McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, 9, 103 – 113.
- McDaniel, M. A., Morgesen, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, 86, 730 – 740.
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality & Social Psychology Bulletin*, 31, 166 – 180.
- Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, 89, 187 – 207.

- Paulhus, D. L. (2006). *Socially desirable responding: A history of measures*. Paper presented at the ETS Mini-conference on Faking in Noncognitive Assessments. Princeton, NJ: ETS.
- Prelec, D. (2006). A Bayesian method for inducing truthful self-assessments. In P. C. Kyllonen (organizer), *Solving the faking problem on noncognitive assessments* (Invited symposium: Psychological Assessment and Evaluation). Athens, Greece: *26th International Congress of Applied Psychology*.
- Prelec, D., & Weaver, R. G. (2006). *Truthful answers are surprisingly common: Experimental tests of Bayesian truth serum*. Paper presented at the ETS Mini-conference on Faking in Noncognitive Assessments. Princeton, NJ: ETS.
- Reeve, C. L., & Hakel, M. D. (2001, June). *Criterion issues and practical considerations concerning noncognitive assessment in graduate admissions* (Bowling Green State University). Symposium conducted at the meeting of Noncognitive Assessments for Graduate Admissions, Graduate Record Examinations Board, Toronto, Ontario.
- Roberts, R. D., Schulze, R., & Kyllonen, P. C. (2006b). *Technical Advisory Committee on Faking on Noncognitive Assessments*. Princeton, NJ: ETS.
- Roberts, R. R., Schulze, R., & Kyllonen, P. C. (2006a). *ETS mini-conference on faking on noncognitive assessments*. Princeton, NJ: ETS.
- Sackett, P. R. (2006). *Faking and coaching effects on noncognitive predictors*. Paper presented at the ETS Mini-conference on Faking in Noncognitive Assessments. Princeton, NJ: ETS.
- Saucier, G. (in press). Measures of the personality factors found recurrently in human lexicons. In G. J. Boyle, G. Matthews, & D. Saklofske (Eds.), *Handbook of personality theory and testing: Vol. 2 – Personality measurement and assessment*. London: Sage.
- Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., Ramsay, L. J., & Yoo, T-Y. (2003). *Journal of Applied Psychology*, 88, 979 – 988.
- Seiler, S. (2007). Individual differences in the willingness to fake on noncognitive measures. In P. Kyllonen (chair), *The faking problem in noncognitive assessment*. Symposium presented at the American Psychological Association Annual Convention, San Francisco, CA.
- Stark, S. (2006). Applying ideal point IRT models to score single stimulus and pairwise preference personality items. In R. D. Roberts, R. Schulze, & P. C. Kyllonen (chairs), *Technical Advisory Committee on Faking on Noncognitive Assessments*. Princeton, NJ: ETS.
- Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., et al. (2000). *Practical intelligence in everyday life*. New York: Cambridge University Press.
- Stricker, L. J., Rock, D. A., & Bennett, R. E. (2001). Sex and ethnic-group differences on accomplishment measures. *Applied Measurement in Education*, 14, 205 – 218.
- The Conference Board, Corporate Voices for Working Families, Partnership for 21st Century Skills, Society for Human Resource Management (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century workforce*. Retrieved March 28, 2007, from: www.21stcenturyskills.org/documents/FINAL_REPORT_PDF9-29-06.pdf
- Tupes, E. C., & Christal, R. E. (1961/1992). Recurrent personality factors based on trait ratings. *Journal of Personality*, 60, 225 – 251.
- Vannelli, J., Kuncel, N. R., & Ones, D. S. (2007, April). A mixed recommendation for letters of recommendation. In N. R. Kuncel (Chair), *Alternative Predictors of Academic Performance: The Glass is Half Empty*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Walpole, M. B., Burton, N. W., Kanyi, K., & Jackenthal, A. (2001). *Selecting Successful Graduate Students In-depth Interviews with GRE Users*. (Graduate Record Examination) A research project funded by the GRE Board Research Committee, the GRE Program, and the Educational Testing Service Research Division.
- Walters, A., Kyllonen, P. C., & Plante, J. (2006). Developing a standardized letter of recommendation. *The Journal of College Admission*, 191, 8 – 17.

Walters, A. M., Kyllonen, P. C., & Plante, J. W. (2003). *Preliminary Research to Develop a Standardized Letter of Recommendation*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Waugh, G. W., & Russell, T. L. (2003). *Scoring both judgment and personality in a situational judgment test*. Paper presented at the 45th Annual Conference of the International Military Testing Association. Pensacola, FL.

White, L. A., Young, M. C., Stark, S., Drasgow, F., & Hunter, A. (2006). New approaches for measuring self-report temperament constructs in “high-stakes” military testing. In P. C. Kyllonen (organizer), *Solving the faking problem on noncognitive assessments* (Invited symposium: Psychological Assessment and Evaluation). Athens, Greece: *26th International Congress of Applied Psychology*.

Zickar, M. (2006). *Using multigroup item-response theory to better understand faking*. Paper presented at the ETS Mini-conference on Faking in Noncognitive Assessments. Princeton, NJ: ETS.

Ziegler, M. (2006). People fake! – So what? In R. D. Roberts, R. Schulze, and P. C. Kyllonen (chairs), *Technical Advisory Committee on Faking on Noncognitive Assessments*. Princeton, NJ: ETS.

Ziegler, M. (2007). People fake! – So what? In P. Kyllonen (chair), *The faking problem in noncognitive assessment*. Symposium presented at the American Psychological Association Annual Convention, San Francisco, CA.

For More Information

To get the most up-to-date information about the ETS® Personal Potential Index, please visit www.ets.org/ppi or contact an ETS representative at ppi@ets.org.



Listening. Learning. Leading.®

www.ets.org