Noncognitive characteristics can be assessed in many different ways, including self-assessments, interviews, and behavioral observations. The most commonly used methods shown in the table below include both well-established and recently developed assessment tools and are organized by source (self or other) and type (ratings or performance).

<table>
<thead>
<tr>
<th>Self</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratings</td>
<td>Performance</td>
</tr>
<tr>
<td>Self-Assessments (Likert Type)</td>
<td>Situational Judgment Tests</td>
</tr>
<tr>
<td>Biodata</td>
<td></td>
</tr>
<tr>
<td>Day Reconstruction Method</td>
<td></td>
</tr>
</tbody>
</table>

Different assessments do not always give the same score on a trait. For example, self-ratings may not highly correlate with actual ability or test performance, although the correlation can be increased by following certain procedures.

It also is not necessarily the case that one method is better than the other. Sometimes, two measures independently predict outcome criteria, each adding variance to the other. For example, it has been found that both self-reported and peer-rated Conscientiousness predicted school performance (controlling for intelligence) independent of one another. Note: For a more detailed description of Conscientiousness and other personality variables, please see Five-Factor Model.

Faking is also an issue in noncognitive assessment, particularly when used in high-stakes applications. The fakeability factor of each assessment method is covered in the following descriptions, as well as psychometric topics related to the prediction of educational outcomes.

**Self-Assessments**

Self-assessments are the most widely used approaches for capturing students’ noncognitive characteristics. Most insights concerning the relationship between noncognitive qualities and educational or work-related outcomes stem from research conducted with questionnaires. Self-assessments usually ask individuals to describe themselves by answering a series of standardized questions. The answer format is usually a Likert-type rating scale (for example, in a range from Strongly Disagree to Strongly Agree) but other formats may also be used (such as Yes-No or open answer). Typically, questions assessing the same construct are aggregated and this aggregated score serves as an indicator of the relevant personality domain.

Self-assessments are a relatively easy, cost-effective, and efficient way of gathering information about the individual. Note: For a more detailed description of Conscientiousness and other personality variables, please see Five-Factor Model.

Faking is also an issue in noncognitive assessment, particularly when used in high-stakes applications. The fakeability factor of each assessment method is covered in the following descriptions, as well as psychometric topics related to the prediction of educational outcomes.

**Note:** For a more detailed description of Conscientiousness and other personality variables, please see Five-Factor Model.

For instance, response scale format influences responses. Respondents also vary in their use of the scale and, in general, there are large cultural effects in response style. For example, young males tend to use extreme answer categories, as do Hispanics.
Respondents can also fake their responses to appear more attractive to a prospective employer or institution, resulting in decreased validity. ETS has conducted several mini-conferences addressing the faking problem and has identified several promising methods for collecting self-assessments, such as giving real-time warnings, using a multidimensional forced-choice format (pitting equally attractive noncognitive factors such as “works hard” and “works well with others” against each other), and using one’s estimates of how others will respond to help control for faking. However, evidence for their effectiveness in controlling faking has not been demonstrated unequivocally.

Self-assessments have been used in many research studies involving noncognitive variables at the Center for Academic and Workforce Readiness and Success.

A selection of peer-reviewed articles and reports that ETS has published on the topic of self-assessments over the past few years follows:


Situational Judgment Test (SJT)

A situational judgment test (SJT) is one in which participants are asked how best to or how they might typically deal with some kind of situation. For example, a situation might be a group project in which one member did not help out. Possible responses might be to:

- talk to the nonparticipating member in private
- talk to the nonparticipating member in front of the group
- let the incident pass without comment

Situations can be described in words or videotaped, and responses can include multiple choice, constructed
response, and ratings (“How good would this response be?”). SJTs can be regarded as fairly simple, economical simulations of job tasks.

SJTs have been shown to predict many different criteria such as college success, army leadership, and managerial performance. Though applications in education have been relatively limited, applying SJTs as a predictor in educational domains has received increased interest.

SJTs may be developed to reflect more subtle and complex judgment processes than are possible with conventional tests. The methodology of SJTs enables the measurement of many relevant attributes of individuals, including leadership, the ability to work with others, achievement orientation, self-reliance, dependability, sociability, agreeableness, social perceptiveness, and conscientiousness. Numerous SJTs, ranging from print-based measures of business analysis and problem solving to video-based measures of communication skills, have been developed.

Research on SJTs has revealed that respondents are able to improve their score in a retest or after coaching, although the improvement may be small. SJTs appear to be less susceptible to faking compared to Likert-type self-assessments, where the improvement due to incentives can be up to a full standard deviation.

Among the following are peer-reviewed articles, reports, and book chapters that ETS has published on the topic of SJTs:


Biodata

Biographical data or “biodata” have been explored for college admissions use in the United States and Chile. Biodata are typically obtained by asking standardized questions about individuals’ past behaviors, activities, or experiences. A sample question could be: “How often in the last two weeks have you eaten fast food?” Respondents are given multiple-choice answer options or are requested to answer in an open format (e.g., frequency).

ETS collects biodata in most of our research studies, including:

- Gender
- Race/Ethnicity
- Age (specific via open response or a range via multiple choice)
- Socio-economic Status (via items that include a range of income, financial aid status, number of books in a home, etc.)
- Grade Point Average (GPA) (via open response and restricted ranges) & Coursework (courses taken, grades, AP® courses, etc)
- SAT®/ACT® Scores
- Citizenship Status
- College Enrollment Status
• College Major
• Attendance
• Languages (first language, language spoken at home, etc)
• Parental Status
• Access to Computers
• Work Status (full-time, part-time, unemployed, hours worked)
• Parental Education
• Level of Schooling (current and aspirations)
• Health Information (weight, height)

We have developed a biodata (documented accomplishments) measure that produced scores for six scales:36, 37

• Academic Achievement
• Leadership
• Practical Language
• Aesthetic Expression
• Science
• Mechanical

For the Leadership category, for example, items were:

• Was on a student-faculty committee in college. Yes/No
• If YES: Position, organization and school?

Measures of biodata have been found to be incrementally valid beyond the SAT test and the Big Five in predicting students' performance in college.38 Obviously, biodata can be faked but there are several ways to minimize faking.39, 40 Asking students to verify with details, for example, can minimize faking.

**Day Reconstruction Method (DRM)**

The day reconstruction method (DRM) measures how people spend their time and how positively or negatively they feel while conducting certain activities.41 Participants list details from several episodes they participated in during the previous day and state when each episode occurred, what they were doing, who they were with, and how they felt. Research has demonstrated that recalling episodes from the previous day is close enough in time for people to accurately recall emotions without bias.42

In our initiative, we have used the DRM in two ways. First, we have used it as a measure of hedonic well-being or the daily experience of positive and negative emotions. For example, we have found that hedonic well-being, is correlated with emotional intelligence.43 We have also used the DRM as part of a time management intervention, and found evidence that suggests it may be useful in improving this skill.44 We believe the act of recording the events of one’s day in detail can lead to improved self-awareness of how time is spent.

**Day Reconstruction:**

• **Lawrenceville Studies:** These studies specifically focused on improving the time management skills of high school students at a private school; the DRM was used as a homework assignment in an effort to increase awareness of how time was spent and how one felt during each activity for each day. A paper currently under peer review has been written on this topic and is available upon request.

• **Army Research Institute (ARI) studies:** The focus of the ARI studies was to assess emotional abilities (Phase I), and their impact on well-being (Phase II), stress (Phase III) and leadership (Phase IV). Two- and four-year college students responded to the DRM during Phase II to assess the time use and emotions associated with activities during the previous day.45

**Other-Ratings**

Other-ratings are assessments in which others (e.g., supervisors, trainers, colleagues, friends, faculty advisors, coaches) rate individuals on various noncognitive qualities. This method has a long history and countless studies have been conducted that employed this method to gather information.46 Other-ratings have an advantage over self-ratings in that they preclude socially desirable responding, although they do permit rating biases. Self- and other-ratings do not always agree,47 but other-ratings are often more predictive of outcomes than self-ratings.48, 49

**Other-Ratings Used at ETS:**

• **ETS® Personal Potential Index/ETS Summer Intern Standardized Letter of Recommendation Study:** In this study, the advisors and faculty members of prospective graduate students rated these students
in the following areas: Knowledge and Creativity, Communication Skills, Teamwork, Resilience, Planning and Organization, and Ethics and Integrity on a scale that ranged between Below Average and Truly Exceptional.50

- **Studies of other-reported noncognitive factors.**
  Because of the validity threat posed by faking in self-assessments, many of the papers described in that section also include other-reports (i.e., some combination of teacher-, parent- and/or peer-reports). A recent study also explored the topic of other-reported SJTs, to present knowledge the first study of its kind.51

**Transcripts**

Transcripts contain information on the courses students have taken, earned credits, grades, and grade point average. As official records, transcript information can be taken as more accurate than self-reports. Transcript data can be standardized and used in validity studies. For example, the U.S. National Center for Educational Statistics supports an ongoing collection of transcripts based on the **National Assessment of Educational Progress (NAEP)** High School Transcript Study, which classifies courses, computes grade point average, and links resulting data to NAEP achievement scores.

**Interviews**

Interviews are the most frequently used method of personnel selection,52 but also are used for school admissions, promotions, scholarships, and other awards. Interviews vary in their content and structure. In a structured interview, questions are prepared before the interview starts. An unstructured interview simply represents a free conversation between an interviewer and interviewee, giving the interviewer the freedom to adaptively or intuitively switch topics. Research has shown that unstructured interviews lack predictive validity53 or show lower predictive validity than structured interviews.54

The best practices for conducting interviews are:55

- High degree of structure
- Selection of questions according to job requirements
- Assessment of aspects that cannot be better assessed with other methods
- Scoring with pre-tested, behavior-anchored rating scales
- Empirical examination of each question
- Rating only after the interview
- Standardized scoring
- Training of interviewers

Structured interviews can be divided into three types:

- Behavioral description interview — involves questions that refer to past behavior in real situations, also referred to as job-related interview56
- Situational interview — uses questions that require interviewees to imagine hypothetical situations (derived from critical incidents) and state how they would act in such situations57
- Multimodal interview — combines the two approaches above and adds unstructured parts to ensure high respondent acceptance58

Analyses of predictive validity of interviews for job performance59, 60, 61, 62 have shown that they are good predictors of job performance, add incremental validity above and beyond general mental ability, and that behavioral description interviews show a higher validity than situational interviews. Interviews are less predictive of academic performance as compared to job-related outcomes. Predictive validity probably also depends on the content of the interview, but the analyses aggregated interviews with different contents.

**Behavioral Observations**

Behavioral observations entail watching observable activities of individuals and keeping records of the relevant activities.63 Records can vary from videos, photographs, and cassette recordings to notes taken by the observer. The general assumption behind this method is that individuals vary in observable behaviors; this variation is stable over time and across different situations, and can be regarded as an indicator of a personality trait.64

One form of behavioral observation often used in selection is the assessment center. Assessment centers can comprise many different methods (including achievement tests), but they feature role play and presentation
A strength of assessment centers for measuring personality is that they are performance-based rather than opinion-based self-assessments. As such, they are less easily faked than are self-assessments. On the other hand, a drawback to assessment centers is that they assess maximum performance, which may not be representative of typical behavior.

Letters of Recommendation

Letters of recommendation can be seen as a more subjective form of other-ratings and have been extensively used in a broad range of situations. They provide stakeholders with detailed information about the applicants’ past performance, with the writer’s opinion about the applicant being expressed in the form of an essay.

In response to a major drawback of letters of recommendation — their nonstandardized format — a more structured system, initially coined the Standardized Letter of Recommendation and now the ETS® Personal Potential Index (ETS® PPI), has been developed. This assessment system prompts faculty members to respond to specific items using a Likert scale, in addition to eliciting comments. It has been used operationally at ETS for selecting summer interns and fellows, as well as through Project 1000 for the selection of graduate student applicants. ETS PPI has been offered as a supplement to GRE® tests beginning in 2009.

Footnotes


