Pearson

Raven's™ Adaptive

User's Guide and Technical Manual

Pearson
TalentLens

Raven's™ Adaptive

User's Guide and Technical Manual

Table of Contents

Tables

Introduction

The Raven's™ Adaptive is designed to measure an individual's ability to learn or to succeed in abstract reasoning. It is widely used in organizations for employee selection and development and in academic settings for educational placement and vocational counseling. The skill assessed with the Raven's™ Adaptive is important in many work settings and is applicable to executive, managerial, supervisory, professional, sales, administrative, and technical roles across most industry sectors.

The Raven's™ Adaptive is an untimed, 15-item computer adaptive test that uses an updated high-quality bank of items to measure the fluid intelligence.

Background

The Raven's Progressive Matrices have been used in many countries for decades as a measure of problem-solving and reasoning ability (Raven, Raven, & Court, 1998a). The various editions of the Raven's Progressive Matrices (standard, advanced and colored) have been studied in more than 48 countries on samples totaling more than 240,000 participants (Brouwers, Van de Vigver, & Van Hemert, 2009; Wongupparaj, Kumari, & Morris; 2015). See Part 2 of the Raven's APM III International Manual for a description of the development and standardization of the Raven's Advanced Progressive Matrices (APM) fixed form (also known as APM short form or APM 2.0) and the subsequent item-banked version designed for use within the domain of work and organizational psychology.

Raven's™ Adaptive Overview

Raven's™ Adaptive is an individually administered, online, adaptive assessment of nonverbal reasoning ability. This test can be administered unsupervised from a remote location or under supervision for greater control over test conditions, examinee identity, and behavior. Each administration of the test includes 15 items that are randomly selected from a pool of 174 items. This ensures no two examinees receive the same test, enables examinees to take the test in an unsupervised setting, and improves test security.

Abstract Reasoning items consist of universal geometric shapes that are recognizable regardless of education level. For each item, an incomplete patterned matrix or series is presented and the examinee must select the response option that correctly completes the matrix or series. In order to select the correct answer, the examinee must detect the principle(s) governing each pattern. Because this test requires minimal verbal instruction and no spoken or written responses, it minimizes the impact that language skills and cultural background may have on the examinee's test performance. The reduced language requirements and cultural load makes it ideal for screening groups of job applicants whose native languages differ or who are dispersed across varying geographies.

What Raven's™ Adaptive Measures

Raven's™ Adaptive measures various aspects of nonverbal reasoning ability. It is an excellent measure of general intelligence, or *g* (Carroll, 1993), which is strongly associated with job performance (Hunt & Madhyastha, 2012; Kuncel, Ones, & Sackett, 2010; Schmidt, 2014). Evident in most abstract reasoning tasks is fluid reasoning, which is the ability to solve novel problems independent of any previously learned knowledge. One hallmark of fluid reasoning is inductive reasoning, which involves recognizing underlying conceptual relations that define how elements of a problem behave together and applying those identified concepts to demonstrate understanding of the relations (Schneider & McGrew, 2018).

Inductive reasoning ability predicts academic and career success in such fields as medicine, engineering, business, research, education, law, law enforcement, and both the hard and social sciences (National Center for O*NET Development, 2018) and is required in varying degrees in occupations such as physician, engineer, researcher, forensic expert, police officer, sociologist, and teacher (National Center for O*NET Development, 2018).

Suggested Applications

Cognitive abilities underlie performance in both work and school. Measures of cognitive ability have been shown to be the single best predictor of educational achievement (Deary & Johnson, 2010; Deary, Strand, Smith, & Fernandes, 2007; Johnson, Deary, & Iacono, 2009; Kaufman, Reynolds, Liu, Kaufman, & McGrew, 2012; Kuncel & Hezlett, 2010; Nelson, Canivez, & Watkins, 2013) and the most effective tool for selecting successful employees (Bertua, Anderson, & Salgado, 2005; Hunt & Madhyastha, 2012; Kuncel & Hezlett, 2010; Kuncel et al., 2010; Lang, Kersting, Hulsheger, & Lang, 2010; Schmidt, 2014; Schmidt & Hunter, 2004).

Results from this test may be used in employment and educational contexts to facilitate pre-employment selection, development, and outplacement or career counseling services. Administrators of this test should have appropriate qualifications in test use and interpretation (e.g., relevant training or credentialing in an occupational or higher education setting).

Selection

Many organizations use testing as a component of the employment selection process to screen out unsuitable applicants and/or to rank order applicants by merit. Using reasoning tests can help employers make more informed employment decisions (Schmidt & Hunter, 2004). Results from this test may be used to rank order applicants or, in combination with other assessment methods, to provide a full profile of an applicant.

Development

Tests can be helpful in better understanding a person's strengths and weaknesses so that appropriate career development goals and activities can be set. People with insight into their own abilities can more easily identify how to capitalize on strengths and minimize the impact of weaknesses. For example, they may be able to make more suitable career choices or to identify projects and tasks that provide opportunities to develop skills that need improvement. Tests also enable organizations to identify skill improvement opportunities for employees to enhance their work effectiveness and career success.

Outplacement and Career Guidance

The results of this test can be helpful in outplacement or career guidance situations (e.g., employees facing job elimination due to redundancy, seeking new opportunities due to changes in life circumstances, or experiencing constrained career advancement opportunities and seeking alternatives). The purpose of assessment is to provide information needed to make realistic occupational decisions that best suit an individual's abilities, needs, and interests. Administrators should evaluate if test results provide greater clarity about an individual's potential to succeed in relevant roles or training. Administrators also should avoid interpreting test scores in isolation and overinterpreting differences between test scores. An examinee's interests, motivations, and circumstances are also important considerations when providing career guidance.

Testing Considerations

Testing Environment

The following are necessary to ensure accurate scores and to maintain the cooperation of the examinees:

- good lighting and ventilation;
- comfortable seating;
- adequate desk or table space;
- space between tables that allows administrators to circulate without disturbing examinees;
- a quiet environment free from distractions and outside interruptions (posting a "Testing in Progress" sign may help);
- examinee's materials, if needed (i.e., calculator, paper, pencil);
- comfortable positioning of the computer screen, keyboard, and mouse; and
- a pleasant and professional attitude on the part of the administrator.

For group sessions, the room should be large enough to provide ample space between tables for examinees to sit comfortably while minimizing conversation and preventing examinees from looking at others' responses. Ideally, there should be at least one empty seat between examinees and if seated at tables, and they should not sit directly across from one another. Leave adequate space between tables to allow administrators to circulate without disturbing the examinees.

Browser Requirements

The online testing platform is reliable and stable. If a computer loses connectivity during the test, all responses are saved. When connectivity is restored and the computer is logged back into the platform, testing resumes at the last unanswered test item.

The following Internet browsers are compatible with the platform:

- Internet Explorer® 8.0 or higher
- Edge®

- Firefox® (latest version, must use auto-update)
- Google Chrome™ (latest version, must use auto-update)
- Safari® (Mac 5.0+)

No additional hardware or software is required. If administrators or examinees encounter technical difficulties, contact TalentLens for assistance.

User Responsibilities

The administrator is responsible for ensuring that test use complies with the code of practice of the testing organization, applicable government regulations, and the recommendations of the test publisher. The administrator should also assume responsibility for ensuring that examinees are informed before the testing session about the nature of the assessment, why the test is being used, the conditions under which they will be tested, and the nature of any feedback they will receive. The administrator should ensure that relevant examinee identifying information is collected and verified (e.g., name) and any information needed for interpretation is available.

Scores are confidential and should be stored in a secure location accessible only to authorized individuals. It is unethical and poor test practice to allow individuals who do not have a legitimate need for the information to access test scores. Store test scores in a locked cabinet or password-protected file that can only be accessed by authorized individuals. Testing material security and protection of copyright are the responsibility of authorized individuals. Authorized individuals and test administrators should avoid disclosure of usernames, passwords, or any other methods that allow access to the test.

Verifying Performance

When initial testing is unsupervised, administrators may wish to verify an examinee's performance by administering the test again in a supervised setting. Each administration of the test includes items that are randomly selected from a larger pool, which ensures that the examinee will not receive the same test.

Scores differ somewhat from one test administration to another. The standard error of measurement (*SEM*) provides an estimate of how much scores can be expected to differ and helps administrators identify pairs of scores that are very unlikely to be attained by the same examinee. When the second score is substantially lower than the initial score, administrators may wish to flag the score report for further consideration. Although the probability of an examinee obtaining a much lower score from a second administration is low, it does occur; particularly when the examinee is experiencing high anxiety or is fatigued from undergoing multiple interview and test procedures.

Test Administration Modifications for Examinees with Disabilities

The U.S. Americans with Disabilities Act (ADA) of 1990 requires an employer or prospective employer to reasonably accommodate the known disability of a qualified applicant, provided such accommodation(s) would not cause an "undue hardship" to the operation of the employer's business. Reasonable accommodations enable examinees with special needs to comfortably take the test. Reasonable accommodations may include, but are not limited to, modification of the assessment format and procedure (e.g., live assistance in which an intermediary reads the test content to a visually impaired examinee and marks the examinee's answers; Society for Industrial and Organizational Psychology, 2003). Consult with your qualified legal advisor or human resource professional for additional guidance on providing appropriate, reasonable accommodations.

Test Administration Guidelines

This test is administered via the TalentLens online testing platform at TalentLens.com. The platform is an Internet-based system designed for administering, scoring and reporting results of professional tests. Administrators of this test should have appropriate qualifications in ability test use (e.g., relevant training or credentialing in an occupational or higher education setting).

The TalentLens online product catalog includes a variety of technical information:

- this manual,
- sample reports,
- responses to frequently asked questions, and
- descriptions of the samples used to generate the norms (i.e., norm composition tables).

Administration Directions

Be thoroughly prepared before administering the test to maximize efficiency. Verify that the organization's TalentLens account provides access to the test. Administrators should be familiar with the administration instructions. Administrators who take the test themselves (complying with all directions) prior to testing others often improve their own familiarity with the test's procedures and ability to anticipate questions or issues that may arise. Administrators should provide examinees with pencils, an eraser, and a sheet of paper to write if needed.

This test can be administered in a supervised or unsupervised setting. The administration procedures for each setting are described below. Scoring is automatic, and score reports are immediately available to test administrators via the online testing platform.

Unsupervised

Verify the email addresses for all examinees. Email or speak to each examinee to provide all of the relevant information (e.g., purpose of test, confidentiality, online administration, if and how feedback is provided). Give remotely located examinees a method to report technical problems (ideally, to their administrator). Ask examinees about any disabilities or special accommodations required. Do not change the standardized administration procedures without seeking advice from an expert, as some adjustments or accommodations may reduce the applicability of normative data to an examinee's results. Contact TalentLens if you are unsure about the impact of a given accommodation that is being considered.

After adding an examinee to the platform, the system generates an email invitation that can be amended before it is sent. The platform administers the test according to the standardized procedures, then scores the test and provides a report immediately. It is strongly recommended to verify unsupervised test results under supervised conditions prior to making final placement decisions.

Supervised

When scheduling any testing session, consider the average length of administration, how many examinees will be tested, the location, and the number of trained administrators necessary. Email or speak to each examinee to provide all of the relevant information (e.g., purpose of test, confidentiality, items or information to bring to the session, online administration, if and how feedback is provided). Ask examinees about any disabilities or special accommodations required. Do not change the standardized administration procedures without seeking advice from an expert, as some adjustments, accommodations, or changes to instructions may reduce the applicability of normative data to an examinee's results. Contact TalentLens if you are unsure about the impact of a given accommodation that is being considered.

Prepare a test log (see Appendix A). This can function as a register and detail any reasonable accommodations made for examinees with disabilities or any unusual occurrences.

Prepare the testing room consistent with the description in the Test Environment section of this manual. Check all computer equipment to ensure that it is in working order and that the examinees have been added to the system. Post a "Testing in Progress" sign. Do not allow examinees who arrive late to join a group session; reschedule testing with those examinees for another time.

It is important that examinees make their best effort. A friendly, purposeful, relaxed approach and nonthreatening tone puts examinees at ease and enables them to do their best. Start the testing session by introducing yourself and your relationship to the organization. Address the following issues.

- purpose of testing
- how results will be used
- who will have access to the results
- if, how, and for how long results will be stored; particularly addressing data security
- what they should do when they have completed the test (e.g., remain seated until instructed)
- what will happen after the testing session

- other practical issues or questions that could arise during the session (e.g., breaks, fire alarms, duration, restroom locations)

Ask examinees to silence and turn off mobile phones and any other portable communication devices (e.g., smart watches, tablets). Handbags, briefcases, and other personal materials on or near the examinee's work surface should be set away from the testing area in a secure location. Any books or notes that examinees have with them must be placed out of reach. Ensure the initial test administration screen is displayed and the examinee is seated. Say,

The on-screen directions will take you through the entire process. First, some identifying information is gathered. Then you will review the test instructions and complete some example items. After reading all instructions and completing the example items, you will be asked if you are ready to begin the test. You will be required to respond Yes or No. Once you click Yes, the test will begin. You can only move forward in the assessment, so you will be unable to access any of the previous screens or items.

After the examinee responds to the necessary items and exits, the system locks, and the examinee cannot return to the test.

If an examinee encounters technical problems during the testing session, the administrator should move him or her to another computer, if possible, and log the examinee back into the system. If another computer is unavailable, or if the technical problems cannot be resolved in this manner, contact TalentLens for assistance.

When testing is complete, thank the examinees for attending. Explain the next steps in the assessment process, then dismiss them. Complete the test log and secure all test materials. Avoid disclosure of test access information, such as usernames and passwords, and ensure that all other necessary data protection procedures are followed.

Interpreting Results

Test scores should be interpreted within the context for which the test is being used. An examinee's scores should be compared to an appropriate normative group (norm group) to obtain an accurate profile of his or her ability. It is also important to integrate test results with other assessment information collected.

Norm Group

The basis for test interpretation relies upon norms. Norms are developed in the following manner. A sample of people is selected by particular characteristics to define a desirable comparison group (i.e., norm group) that is representative of the population with whom the test will be used. The norm group is tested, and their scores (i.e., norm data) are transformed from a frequency distribution into a standard scale with known properties (i.e., norms). An examinee's test score is compared with the norms to determine how the examinee performed relative to the norm group. Two types of norm groups can be used to facilitate test interpretation: the published norm group, and local norm groups, described below.

Published Norm Group

The published norm group for this test was selected from among working adults. Further details about the norm group appear in the Norms Composition table, which is available on the TalentLens online product catalog in the technical information section.

Organizations can use the published norms to benchmark the performance of their employees or job applicants relative to that of other working adults. For career guidance, the published norm group may be most appropriate as these enable the administrator to benchmark the examinee's skills more broadly.

Local Norm Groups

Organizations can also use a different comparison group to create their own norms. This allows the organization to judge the examinee's performance in relation to others in similar roles. When an organization tests many examinees, it may be useful to create a local norm reflecting the performance of applicants or incumbents in the organization.

A local norm group must be based on a sufficiently large group of people (ideally, at least 150) that is representative of the examinees with whom the test will be used. If there is insufficient data to create a local norm, or the group is unrepresentative in some way (e.g., available scores come from recently graduated recruits and you wish to assess more experienced managers), it may be preferable to use the more general published norm group of working adults. Local norm groups should be appropriate for the purposes of test use. If possible, the comparison group should have taken the test for a similar purpose (e.g., selection, development).

Local norm groups ideally should have applied for roles at a similar level, although industry sector and job type are also important. Where the job level is not clear, typical educational background may indicate how the local norm group should be composed. Where possible, the norms should reflect the composition of the potential examinee group with respect to demographic characteristics (e.g., age, sex, race/ethnicity). If you would like to create a local norm for comparison purposes, please contact TalentLens.

Understanding Score Reports

Scoring is automatic and the examinee's Profile Report is usually available instantly. A link to the report is available in the administrator's account in the online testing platform. Adobe® Acrobat Reader® is required to view and download the report. The administrator may view, print, or save the examinee's report.

Test scores should be compared to appropriate norm groups and interpreted considering the purpose of testing and other related information. The norm group should be described to give a clearer picture of the examinee's performance in context, as that group is the backdrop against which the examinee's scores are compared. It is important to connect the scores to the testing purpose; that is, what scores mean in terms of a career or a specific job. It is also necessary to integrate test results with other information that is collected (e.g., other test scores, interviews, work sample).

Because there are many different forms of the test, each with minor variations in difficulty, scoring algorithms based on item response theory (IRT) are used. The algorithms take into account the exact difficulty level of the items the examinee completed and produce a "theta" score.

The Profile Report includes the examinee's scores: percentile rank and several other norm-referenced scores. Raw scores should never be used to make decisions because item-banked tests take into account the difficulty of each item. Instead, the percentile rank or one of the standard scores that account for the minor differences in difficulty between items in different administrations should be used. The advantage of a standard score is that it has a consistent meaning across different tests and comparison groups. Percentile ranks and standard scores describe the examinee's level of performance.

Theta scores have been omitted from the computer-generated reports because many test users are unfamiliar with them. Instead, traditional score types are reported for easier interpretation. Administrators can extract theta score data free of charge, however, from the online testing platform. Theta scores range from –3.00 to +3.00. A theta score provides an estimate of an examinee's ability level that takes into account, among other factors, the difficulty level of the items answered correctly. Therefore, two examinees that answer the same number of items correctly may have different theta/percentile ranks/*T* scores if the questions they answered varied in difficulty. This method allows harder questions to be weighted more heavily when judging performance.

Percentile Rank

The percentile rank indicates a score's standing relative to those scores obtained by the norm group. They reflect the percentage of examinees in the norm group scoring equal to or below that score. For example, if a score is at the 75th percentile of a given norm group, it means that the examinee scored higher than or equal to 75% of the norm group. A score above the 50th percentile is considered above average in relation to the norm group.

Although easy to understand and useful for explaining an examinee's performance relative to that of others in the norm group, percentile ranks have various limitations. Percentile ranks do not have equal intervals. In normal distributions, which are commonly seen in cognitive ability data, percentile ranks tend to cluster near the median (the 50th percentile). Consequently, for scores within the average range, a change of 1 or 2 points may produce a large change in the percentile rank. For scores that are more extreme, a change of 1 or 2 points is not likely to produce a sizable change in their percentile ranks. For this reason, it is not appropriate to sum or correlate percentile ranks with other scores.

T Score

The *T*-score scale has an average score of 50 and a standard deviation (*SD*) of 10. Higher scores indicate better performance. When scores are normally distributed, 68% of examinees score between *T* scores of 40 and 60.

The advantage of *T* scores is that they represent an even scale—that is, the difference between scores of 70 and 80 is the same as the difference between scores of 45 and 55. It is possible to add and subtract *T* scores and to correlate them with other measures.

T scores provide a good level of differentiation for ability tests because there are enough points on the scale to represent all the different score levels. Generally, *T* scores should not be used in feedback to test takers, as they can be difficult to comprehend without some understanding of statistics.

Sten Score

Sten scores have a 10-point-scale with a mean of 5.5 and an *SD* of 2. Higher scores indicate better performance. A sten score of 10 is in the top 2.3% of scores in the norm group. Like *T* scores, sten scores are an even scale, but the smaller range often is easier to understand.

Stanine Score

Stanine scores have a 9-point scale with a mean of 5 and an *SD* of 2. Stanine scores are commonly used in feedback. Higher scores indicate better performance. A stanine score of 9 is in the top 4% of scores in the norm group. Like *T* scores and sten scores, they are an even scale. Similar to sten scores, the smaller range often is easier to understand.

Accuracy of Test Scores

Scores obtained on any test can be considered only an estimate of the examinee's true score. This is because no test is perfectly accurate (without error). The *SEM* indicates the amount of error to be expected in an examinee's score. The amount of error can be expressed as range of raw score points (i.e., number correct), standardized scale points, or percentile points. In any case, an examinee's true score lies within that range. The *SEM* is described in more detail in the Evidence of Reliability section in this manual.

Limitations of Test Scores

Interpret test scores carefully. Errors may arise during test administration. Scores also can be affected by an examinee's state of mind (e.g., feeling anxious or ill). Examinees with a disability or who speak English as a second language may be disadvantaged due to the test format. For these reasons, interpret them with caution. This test is intended to be used alongside other assessment methods.

On occasion, test scores may contradict alternative information about an examinee. When this happens, the administrator should work with the examinee to explore the information and discover possible causes for these anomalies.

Interpret results of unsupervised test sessions with caution, unless there is certainty that the test was completed without assistance. Results from unsupervised test sessions can be verified in a number of ways. For example, a smaller group of examinees may be retested in a supervised setting during a later stage of the decision-making process, or information can be obtained from other sources (e.g., structured interviews or assessment center exercises measuring the same abilities).

Using Test Scores for Employee Selection

Cognitive abilities underlie performance in both work and school. Measures of cognitive ability have been shown to be the single best predictor of educational achievement (Deary & Johnson, 2010; Deary et al., 2007; Johnson et al., 2009; Kaufman et al., 2012; Nelson et al., 2013) and the most effective tool for selecting successful employees (Hunt & Madhyastha, 2012; Kuncel et al., 2010; Schmidt, 2014; Schmidt & Hunter, 2004). Many organizations use cognitive tests to screen out unsuitable applicants, to rank order applicants by merit, and/or to complement other selection information to find the most suitable applicant. Results from this test may be used to compare applicants or in combination with other assessment methods to gain a more comprehensive understanding of an applicant.

Before using the test as part of the selection process, organizations should ensure that the test is relevant to the role for which they are recruiting. Using inappropriate tests or relevant tests in an inappropriate manner can result in poor and unfair decisions. Job analysis is the process of breaking down a job into its tasks, requirements, and performance criteria so that recruiters have a clear understanding of what a job entails. Formal job analysis methods are most effective (e.g., questionnaires, critical incident analysis); but, at a minimum, there should be a discussion with people who know the job well. It is advantageous to talk to both managers and job incumbents, as they may have different perspectives on the role. Other informants (e.g., customers, trainers, reports) may also be helpful. The collected information should be used to write a job description and a person specification. A job description lists components of the job, duties or tasks, responsibilities, and the required standards of performance. A person specification lays out the personal characteristics necessary to do the job (e.g., specific skills and abilities), and it frequently includes a competency profile as well.

The job description and person specification should be used to select the type of assessment most relevant to the role. The administrator should review the test manual to ensure that the test is relevant in terms of its level and the group to be tested. The applicability of norms and the test qualities (e.g., reliability, validity, and group comparisons) should also be considered. The standards of the assessment should not be higher than what the job requires. Normed scores are used to understand the ability level of a particular examinee or to compare examinees for short-listing purposes. However, a short list should not be based on a single measure, as this only reflects a single aspect of performance.

Results from this test enable you to make more informed decisions about an applicant's ability and minimize the chance of a bad hiring decision because of poor recruitment. A good understanding of the job role and careful selection of tests and norm groups ensures sound evidence for the decision to use a test. We recommend documenting this process, as the organization could be required to show that any assessments used were carefully chosen and are relevant, if legally challenged. If you require assistance to determine the appropriateness of this test for selection purposes, please contact TalentLens.

Using Raven's™ Adaptive for Selection

Raven's™ Adaptive measures nonverbal reasoning ability; in particular, inductive reasoning. Inductive reasoning ability was rated as important or higher for more than 80% of the over 900 occupations in the O*NET occupational information database, and for 172 of these occupations it was rated as very or extremely important (National Center for O*NET Development, 2018).

Abstract Reasoning results are used to predict success in positions that require effective problem-solving and decision-making. In particular, results from this test will be relevant when the job or curriculum requires the ability to perceive relationships among things rather than among words or numbers, such as in mathematics, computer programming, drafting, and automobile repair.

Legal Considerations

The U.S. has governmental and professional regulations that cover all personnel selection procedures. Administrators may wish to consult the *Standards for Educational and Psychological Testing* (*Standards*; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014); the *Principles for the Validation and Use of Personnel Selection Procedures* (*Principles*; Society for Industrial and Organizational Psychology, 2003; and the federal *Uniform Guidelines on Employee Selection Procedures* (*Guidelines*; U.S. Equal Employment Opportunity Commission [EEOC], 1978). For an overview of the statutes and types of legal proceedings that influence an organization's equal employment opportunity obligations, the administrator is referred to Cascio and Aguinis (2005) or the *U.S. Department of Labor's Testing and Assessment: An Employer's Guide to Good Practices* (1999).

Fairness in Selection Testing

In any assessment process, it is important to consider issues of fairness and equality of opportunity for both legal and ethical reasons. Fair employment regulations differ across countries, and their interpretation is continuously subject to changes in the legal, social, and political environments. Consult with qualified legal advisors and human resource professionals if you have any questions.

According to the *Guidelines*, adverse impact typically is indicated when the selection rate for one group is less than 80% (or 4 out of 5) of another group (EEOC, 1978). There are well-documented and established age, sex, and racial/ethnic differences on measures of general intellectual ability and specific cognitive abilities (Colman, 2016; Kaufman, Raiford, & Coalson, 2016; Linn & Peterson, 1985; Lipnicki et al., 2017; Maeda & Yoon, 2013; Rushton & Ankney, 2009; Voyer, Voyer & Saint-Aubin, 2017; Wechsler, 2008).

It is essential that administrators are aware of any group differences. The *Guidelines* indicate that a test with adverse impact can be used for selection, but the testing organization must demonstrate that the selection test is job-related, predicts performance, and is consistent with business necessity (EEOC, 1978). A local validation study, in which scores on this test are correlated with indicators of on-the-job performance, provides evidence to support the use of the test in a particular job context. A local study that demonstrates that this test is equally predictive for protected subgroups, as outlined in the *Guidelines*, helps to establish test fairness. Additional guidance on monitoring the selection system for fairness follows.

Monitoring the Selection System

An organization's ability to evaluate selection strategies and to implement fair employment practices depends on its awareness of applicants' and incumbents' demographic characteristics. Monitoring these characteristics and accumulating test score data are necessary for establishing legal defensibility of a selection system. The most effective use of this test is with a local norms database that is regularly updated and monitored incorporating the following best practices over time.

- The hiring organization should ensure that its selection process is clearly job related and focuses on characteristics that are important to job success. At least once every 5 years, conduct a job analysis of the position for which this test is being administered. A job analysis will help determine if the job has changed in a way that requires adjustments to the selection system.
- Periodically (e.g., once every 5 years) reassess the criterion-related validity of the selection system through local validation studies.
- Carefully monitor test scores for evidence of adverse impact. Evaluate adverse impact by comparing the selection rates for individuals from EEOC protected subgroups (e.g., gender or ethnicity) with selection rates of historically advantaged groups. Information needed to facilitate these analyses includes applicant demographics (e.g., voluntary information on gender, race/ethnicity, and age), assessment scores, and employment status (e.g., hired/not hired).
- Periodically re-examine cut scores in light of recent validity results, adverse impact, market data, and other factors (e.g., projected workload), and make adjustments as necessary.
- When sufficient samples of employees and examinees have been obtained (e.g., > 25 per demographic group), analyze the information collected to see if the selection procedure predicts equally for the majority group and EEOC protected groups. The Cleary model (1968) is the most commonly-used approach to evaluate the fairness of selection tools (Aguinas & Smith, 2007; Guion, 2011; Meade & Fetzer, 2009). This model utilizes regression analysis to determine if a test demonstrates differential validity or prediction among subgroups of applicants. The regression approach has been used systemically, is the accepted approach to test for bias, and is supported by the *Standards*, *Principles*, and *Guidelines*. However, the Cleary model has been criticized for its limitations, and alternatives to evaluating test fairness and interpreting results have been proposed (see Aguinas & Smith, 2007; Meade & Fetzer, 2009; and Meade & Tonidandel, 2010 for further discussion and proposed alternatives).

If you require assistance, please contact TalentLens.

Development and Standardization

Rationale and Revision Goals

Development of the Raven's™ Adaptivewas undertaken with the goals to (a) measure the same abilities measured by the prior edition, (b) incorporate significant new features and innovations requested by customers, and (c) maintain the same tradition of quality that established the Raven's™ Adaptive as a market-leading ability appraisal over the past several decades. Therefore, the Raven's™ Adaptive:

- contains updated, contemporary content;
- is brief and makes efficient use of time;
- does not rely on speeded performance;
- requires administration of fewer items relative to the prior edition's retained tests;
- uses modern test assembly and scoring methods based on item response theory (IRT);
- provides a positive examinee experience because it adapts to the examinee's ability level;
- provides strong item security;
- offers many secure, equivalent forms to facilitate both supervised and unsupervised screening of large groups of examinees;
- allows flexible, targeted assessment of a variety of abilities relevant to job performance;
- maintains sufficient reliability;
- has strong international face validity and item applicability;
- is available in multiple languages (beginning with French, (France dialect), followed shortly by English, United States [U.S.] and United Kingdom [U.K.] dialects, French (French Canadian dialect), Dutch, and Spanish (Spain and Latin American dialects);
- provides high levels of score precision for examinees at all skill levels;
- provides state of the art, computer generated reporting; and
- offers the option to obtain a report in languages that differ from the administration language.

Test Development

The main goal of Raven's™ Adaptive development was to create a measure that incorporates sophisticated test construction and statistical analysis techniques to better meet customer needs. Extensive market research was done to gauge customer requirements. This market research was used to determine the development goals.

Test specifications outlined measurement goals, including response options and scoring. The test was developed with an item bank delivery system, whereby each examinee receives a test form composed of a set of items drawn randomly from a large bank. This method greatly reduces the chance of any two examinees receiving the exact same set of items. This means that if an examinee is able to access the test in advance of their testing session (a risk when the test is used unsupervised), the examinee is highly unlikely to receive the same items during a testing session, therefore protecting the integrity of the test.

Item Development

A large item bank is required to support Internet delivery of a great number of parallel forms. Parallel forms are randomly generated, which enables examinees to complete the test unsupervised. The development of the item bank occurred in three main stages described in this section: item writing, item pilots, and psychometric analyses. Item bank development was guided to ensure various parameters were met (e.g., the item-banked test is reliably measuring the intended skills, content is relevant for a global market, content is appropriate for all demographic groups).

Item Writing

The development team reviewed and directed the revision of items. Experienced item writers from Cambridge University's Psychometrics Centre were hired to write the items.

Items were written to be highly relevant to nonverbal inductive reasoning. The classic Raven's™ Progressive Matrices item format was chosen as the format to be used for creating items for the bank. This test format is widely regarded as one of the most robust measures of nonverbal inductive reasoning. The response format was multiple choice. A total of 280 new items were developed for standardization.

Item Pilot

As well as assessing the quality of the items, the goal of the pilot was to link the large set of items together to form an item bank. To accomplish this, anchor items are necessary. Twenty eight anchor items spanning the full range of difficulty were selected from Raven's Standard Progressive Matrices (Raven, Raven, & Court, 2000).

A test item pilot was conducted to determine which of the 280 developed items would be retained in the final test item bank. A total of 10 forms of the test were created for this pilot. Each of the 10 test forms included the 28 anchor items and 28 trial items.

Examinees were recruited through an online research site where participants can complete tests to gain experience with psychometric assessments. Each examinee completed 1 of the 10 fixed-length 56-item forms, rather than an adaptive test, which is the standard approach to developing adaptive test banks (Thompson & Weiss, 2011). This approach is most suitable to allow item parameters to be generated and data to be gathered on a preferred maximum number of items. The test was delivered via the Cambridge University online testing platform Concerto. All testing sessions were unsupervised.

Cases were removed that did not conform to data quality standards. A total of 4,461 participants took part in the study, with final samples sizes after data cleaning ranging per form from 171 to 1,056.

Psychometric Analyses

Item-level analyses were completed to determine which items would be retained in the item bank. A classical test theory (CTT) item analysis was completed first followed by an item response theory (IRT) analysis.

Analysis 1

Each form was analyzed separately. Item difficulty and discrimination were examined and a distractor analysis was performed for each of the 280 items being considered for the final item bank. Scoring keys were verified, and items that were not performing well were removed from the item pool. This included items that did not differentiate well between high and low performers on the test as a whole.

Analysis 2

Specialized statistical analysis software was used to estimate item parameters for the Rasch model and to evaluate their fit. The one-parameter model was chosen as the most appropriate for these data, with each item in the bank assigned difficulty (b) parameters. A total of 112 items were of high enough quality for inclusion in the operational item bank.

This item bank was supplemented with 62 items from the Pearson TalentLens Raven's Advanced Progressive Matrices-III item bank, which were updated to follow the same revised design and color format as the new item bank. A study was conducted to calibrate these 62 items with the new 112 items. Participants (*n*=119) were recruited through an online research site where individuals can select and complete tasks in exchange for payment. Each participant completed a form of 24 items, composed of a combination of Raven's items and new items. Statistical analysis was then conducted to generate item parameters for the 62 Raven's original items, calibrated alongside the 112 new items. This resulted in a final item bank size of 174 items.

Standardization/Normative Group Study

The standardization study focused on collecting and deriving norms and providing evidence of reliability and validity for the final test. Examinees were recruited through an online research site where participants can select and complete tasks in exchange for payment. Each examinee completed an adaptive, fixed-length 15-item form. The test was delivered via the TalentLens online testing platform. All testing sessions were unsupervised. A form completion time limit of 60 minutes was applied. This generous time limit allowed for the vast majority of examinees to complete the entire form.

Cases were removed from the data set when there was evidence that examinees had not approached the test seriously. This included examinees with very low scores and those who completed the test in less than 2 1/2 minutes (i.e., an average of less than 10 seconds per item). Cleaned data were available for most examinees (e.g., 378 for the French, France dialect sample, 768 for the English, U.S. dialect sample, and somewhat fewer for the samples for other languages).

Item Bank Configuration and Computer Adaptive Testing (CAT) Design

The operational, final test is designed so that each examinee receives a test form with 15 items, generated randomly by the online software from the final bank based on the examinee's response to each item.

Scoring is based on item-response theory (IRT), and each item is tagged with an IRT difficulty (*b*) parameter that specifies its difficulty level. This is factored into the scoring algorithm that produces the overall theta score on the test. This methodology adjusts for minor differences in item difficulty to ensure reported scores are on the same scale regardless of the exact set of items administered. Raven's™ Adaptive uses a computer adaptive functionality where 15 test items will be administered randomly based on the examinee's ability level. Test data are captured and scored digitally, and computer-generated interpretive reports are instantly available to the administrator.

Items are presented using an algorithm based on the examinee's ability estimate gained from the items completed previously in the test. Nothing is known about the examinee prior to administration of the first item, so the algorithm is started by selecting an item of medium, or medium-easy, difficulty (*b*-parameter 0, which could be a negative or positive value) as the first item as there is a direct relationship between the item's *b* value and the examinee's ability score (i.e., theta score). Items are presented to examinees based on their ability score. At the start of the test, the ability is assumed to be average, at the 0 level, therefore an item will be presented with a corresponding difficulty level (*b* value) near 0.

The next item selected to be administered depends on the examinee's response to the current item being administered. For example, if an examinee responds correctly to an item of intermediate difficulty, the next presented item is more difficult. Based on the examinee's response to each item (correct or incorrect), the ability estimate is modified. The subsequent items that are presented have difficulty values that corresponds to the modified, current ability estimate. For example, if the ability score increases to 0.8, then the next item presented has a difficulty value of approximately 0.8 because it comes from the pool of 5 items closest to 0.8.

A final ability score is calculated for each examinee at the conclusion of the test. A response is required for each of the 15 items or a score cannot be calculated. The examinee may not go back and revisit a previous item once a response has been entered and they have moved forward to the next item.

It is difficult to calculate the exact number of unique forms that can be generated from the item bank because of the different test specifications that apply. With 15 items drawn randomly from the bank of 174, even conservative estimates suggest that many millions of test forms can be created. This makes it very unlikely that any two examinees will receive the same test form.

Evidence Supporting Reliability

The reliability of a test score refers to its accuracy, consistency, and stability across situations (Sattler, 2008; Urbina, 2014). Classical test theory posits that a test score is an approximation of an examinee's hypothetical true score, that is, the score he or she would receive if the test were perfectly reliable. The difference between the hypothetical true score and the examinee's obtained test score is measurement error. A reliable test has relatively small amounts of measurement error and produces consistent measurement results within one administration and on different occasions. Reliability should always be considered when interpreting obtained test scores and differences between an examinee's test scores on multiple occasions.

Reliability for the Raven's™ Adaptive tests was calculated using IRT-based marginal reliability (Dimitrov, 2003) instead of traditional internal consistency reliability methods (e.g., coefficient alpha, split-half reliability) because each examinee in the normative sample was administered a subset of items from the item bank. IRT-based marginal reliabilities agree closely with internal consistency reliability measures such as coefficient alpha and split-half (Dimitrov, 2003; Wainer & Thissen, 1996); they estimate the consistency of scores that would be obtained on different but parallel sets of items.

A Monte Carlo simulation was conducted to estimate the marginal reliability expected across 2,000 15-item forms randomly assembled from the item bank. For all languages, the mean of the marginal reliability coefficients is .70, which is in the acceptable range. The *SD*s of the 2,000 marginal reliability coefficients are .01. Relatively small *SD*s of marginal reliability coefficients suggest that the estimated marginal reliabilities are very stable across all randomly assembled forms. Hence, the reliability has a very small chance of being lower than .70.

The standard error of measurement (*SEM*) provides an estimate of the amount of error in an examinee's observed test score. The *SEM* is inversely related to the score's reliability. The *SEM* is the *SD* of the measurement error distribution. Table 1 provides the *SEMs* for the *T*, sten, and stanine scores.

Table 1. Standard Errors of Measurement

Score		
T	**Sten**	***Stanine***
5.48	1.10	1.10

The *SEM* is used to calculate confidence intervals, or bands of scores around observed scores, in which true scores are likely to fall. Confidence intervals express test score precision and serve as reminders that measurement error is inherent in all test scores and that observed test scores are only estimates of true ability. Confidence intervals can be used to report an examinee's score as an interval that is likely to contain the true score.

Evidence Supporting Validity

Validity, as described in the Standards (AERA et al., 2014), “is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use” (p. 14). As a result, evaluating a test’s validity requires judgment by the administrator. Three broad categories of evidence are typically recognized for examining the issue of test validity: content validity, criterion-related validity, and construct validity.

Evidence of Content Validity

In an employment setting, evidence of content validity exists when an assessment includes a representative sample of tasks, behaviors, knowledge, skills, abilities, or other characteristics necessary to perform the job. Evidence of content validity is usually gathered through job analysis. The APM has been widely used for decades as a measure of deductive ability—“the ability to evolve high-level constructs which make it easier to think about complex situations and events” (Raven, Raven, & Court, 1998a, p. G8). In an extensive analysis of the cognitive processes that distinguish between higher-scoring and lower-scoring examinees on the Standard Progressive Matrices (SPM) and the APM, Carpenter, Just, and Shell (1990) described the Raven’s Progressive Matrices as “a classic test of analytic intelligence … the ability to reason and solve problems involving new information, without relying extensively on an explicit base of declarative knowledge derived from either schooling or previous experience” (p. 404). In an employment setting, evidence of the content-related validity of the APM should be established by demonstrating that the jobs for which the APM is to be used require the problem solving skills measured by the assessment.

Evidence of Construct Validity

Evidence of construct validity indicates the extent to which the test measures the theoretical construct or trait it is designed to measure. Construct validity can be demonstrated through many types of evidence. If the evidence involves correlations with other measures, the pattern of correlations should reflect the degree of similarity between the two measures. For the following construct validity study, participants were recruited online, compensated for their time, and asked to complete a series of assessments in an online, unsupervised setting.

Correlations with DAT Abstract Reasoning

A total of 104 English speaking examinees in the U.S. completed the 15 items adaptive Raven’s™ Adaptive test along with the 30 item fixed form DAT Abstract Reasoning test (Pearson, 1991). It was anticipated that this correlation would be high, as the two tests measure the same construct. However, since the format of the two tests differed (i.e., one was adaptive and the other was not), and therefore there was no item parameter for theDAT Abstract Reasoning, the correlation between the raw total of the fixed form and the theta score of the adaptive form was calculated and may be somewhat attenuated.

This correlation was high (.64) despite the difference in test format. This result provides good evidence of convergent validity for the scores of these two tests, indicates they are measuring the same construct, and provides construct validation against the previous version of the test.

Correlations with the GAT Abstract

A total of 84 French-speaking examinees in France completed Raven's™ Adaptive along with the Abstract Reasoning test of the General Aptitudes Test ([GAT]; Smith & Whetton, 2011). The GAT Abstract items are similar to those of the Raven's™ Adaptive test. It was anticipated that this correlation would be high because the two tests measure the same construct.

As expected, this correlation was high (.61). This result provides good evidence of convergent validity for the scores of these two tests and indicates they are measuring the same construct.

Correlations with Raven's Advanced Progressive Matrices

A total of 89 French speaking examinees in France completed the 15-item adaptive Raven's™ Adaptive test along with a short form of Raven's Advanced Progressive Matrices (APM; Raven, 1962). It was anticipated that this correlation would be high because the two tests measure the same construct. However, because the format of the two tests differed (i.e Raven's™ Adaptive was adaptive and APM was not), and therefore there was no item parameter for APM, the correlation between the raw total of APM and the theta score of Raven's™ Adaptive was calculated and may be somewhat attenuated.

This correlation was high (.73) despite the difference in test format. This result indicates the two tests are measuring the same construct, and is strong evidence of convergent validity.

Evidence of Convergent Validity

Evidence of convergent validity, a sub-type of construct validity, is provided when scores on an assessment relate to scores on other assessments that claim to measure similar traits or constructs. Years of previous studies on the APM support its convergent validity (Raven, Raven, & Court, 1998b). In a sample of 149 college applicants, APM scores correlated .56 with math scores on the American College Test (Koenig, Frey, & Detterman, 2007). Furthermore, in a study using 104 university students, Frey and Detterman (2004) reported that scores from the APM correlated .48 with scores on the Scholastic Assessment Test (SAT). Evidence of convergent validity for the latest item-banked version of the APM is supported by two findings. First, in the standardization sample of 929 individuals, scores on the item-banked APM correlated .98 with scores on the previous APM. Second, in a subset of 41 individuals from the standardization sample, the revised APM scores correlated .54 with scores on the Watson-Glaser Critical Thinking Appraisal®—Short Form. Detailed evidence regarding the validity of the Watson-Glaser as a measure of critical thinking and reasoning appears in the Watson-Glaser Short Form Manual (Watson & Glaser, 2006)

Evidence of Criterion-Related Validity

Tests are primarily used to predict some future criteria, for example, potential to succeed in a professional role or to benefit from training and development. Examining criterion-related validity informs an administrator's judgments about the potential validity of such predictions that high-scoring examinees will be successful on some criterion of interest. One can statistically determine how much confidence may be placed in using test scores to predict success on the criterion of interest by collecting test scores and criterion scores.

Criterion scores may include job performance ratings, grades in a training course, or scores on other tests. Typically, the correlations of test scores with criterion scores are examined to provide evidence of criterion--related validity.

Cronbach (1970) characterized criterion-related validity coefficients of .30 or better as having "definite practical value." The U.S. Department of Labor (1999) provides the following general guidelines for interpreting validity coefficients: values above .35 are considered "very beneficial," .21–.35 are considered "likely to be useful," .11–.20 it "depends on [the] circumstances," and below .11 are considered "unlikely to be useful." It is important to point out that even relatively low correlations (e.g., .20) may justify the use of a test in a selection program (Urbina, 2014).

Another frequent method of establishing criterion-related validity is to compare test scores across groups of interest that represent levels or aspects of other relevant variables (e.g., education level) and examine the effect size of the standard difference across groups. The term *standard difference* refers to Cohen's *d*. Where data on differences are presented, the effect sizes refers to Cohen's *d*, or the standard difference. Cohen's original suggestions for effect size descriptions indicated that .20 is characterized as small, .50 as moderate, and .80 as large (Cohen, 1988, 1992). Although these ranges do not fully describe all aspects of effect size interpretation, for the purposes of simplicity, values for Cohen's *d* that range from .20 to .49 are reported as small effect sizes. Values that range from .50 to .79 are reported as moderate effect sizes, and values of .80 or greater are reported as large effect sizes.

The practical value of the test depends not only on the validity, but also other factors, such as the base rate for success (i.e., the proportion of people who would be successful in the absence of any selection procedure). If the base rate for success is low (i.e., few people would be successful on the job), tests of low validity can have considerable utility or value. When the base rate is high (i.e., selected at random, most people would succeed on the job), even highly valid tests may not contribute significantly to the selection process. The selection ratio (i.e., percentage of applicants to be selected) also affects the usefulness of a test.

There is abundant evidence that measures of general mental ability, such as the APM, are significant predictors of overall performance across jobs. For example, in its publication on the Principles for the Validation and Use of Personnel Selection Procedures (2003) it is established that validity generalization is well-established for cognitive ability tests. Schmidt & Hunter (2004) provide evidence that general mental ability "predicts both occupational level attained and performance within one's chosen occupation and does so better than any other ability, trait, or disposition and better than job experience" (p. 162). Prien, Schippmann, and Prien (2003) observe that decades of research "present incontrovertible evidence supporting the use of cognitive ability across situations and occupations with varying job requirements" (p. 55). In addition, many other studies provide evidence of the relationship between general mental ability and job performance (e.g., Kolz, McFarland, & Silverman, 1998; Kuncel, Hezlett, & Ones, 2004; Ree & Carretta, 1998; Salgado, et al., 2003; Schmidt & Hunter, 1998; Schmidt & Hunter, 2004).

In addition to inferences based on validity generalization, studies using the APM in the past 70 years provide evidence of its criterion-related validity. For example, in a validation study of assessment centres, Chan (1996) reported that scores on the Raven's Progressive Matrices correlated with ratings of participants on "initiative/creativity" ($r = .28$, $p < .05$). Another group of researchers (Gonzalez, Thomas, & Vanyukov, 2005) reported a positive relationship between scores on the Raven's APM and performance in decision-making tasks. Fay and Frese (2001) found that APM scores were "consistently and positively associated with an increase of personal initiative over time" (p. 120). Recently, Pearson (2010) conducted a study of 106 internal applicants for management positions in which APM scores were positively correlated with trained assessor ratings of "thinking, influencing, and achieving." In addition, manager applicants scoring in the top 30% of APM scores were two to three times more likely to receive above average ratings for the "Case Study/Presentation Exercise", "Thinking Ability", and "Influencing Ability" than applicants in the bottom 30% of APM scores. In addition, the APM Manual and Occupational User's Guide (Raven, 1994; Raven, Raven, & Court, 1998b) provide further information indicating that the APM predicts the ability to attain and retain jobs that require high levels of general mental ability.

Additional discussion of criterion validity for nonverbal inductive reasoning tests and the constructs they measure, as well as new evidence of criterion validity for this edition, appears in the sections that follow.

Relations with Career and Academic Success Variables

In an applied setting, an important indicator of criterion-related validity is whether test scores relate to important organizational outcomes and can help managers identify qualified candidates. Abstract reasoning tests and the constructs they measure (e.g., inductive reasoning, fluid intelligence) have been found to have high *g* loadings (Sattler, 2008) and are associated with academic and professional success. For example, abstract reasoning was significantly associated with academic performance in the first academic year for undergraduate students in Nursing, Social Communication, Law, Engineering, Business, and Medicine programs (Corengia, Pita, Mesurado, & Centeno, 2013). Additionally, inductive reasoning is positively correlated with grade point average (Diaz-Morales & Escribano, 2013), and shares a moderate to high correlation with academic achievement scores (Wechsler, 2008). Fluid reasoning, the broader ability that encompasses inductive reasoning and involves the ability to flexibly solve novel problems (Schneider & McGrew, 2018), shows a strong association with psychosocial adaptation (Huepe et al., 2011) and has been found to predict earnings (Lindqvist & Vestman, 2011).

Education Level Comparisons

Table 2 provides comparisons between groups of examinees with various education levels in the French, France dialect normative sample and the English, U.S. dialect sample. Each group is compared with the group that has the highest education level available for that particular sample. Larger effect sizes (represented by Cohen's *d*) indicate larger differences between groups.

The results in Table 2 indicate Raven's™ Adaptive theta scores increase with education level in the French sample. The mean for the 2 years college or less group is lower than the mean for the 3 years of college or more group, and the effect size is small. For each sample, the group with a higher education level produced higher scores. For the U.S. sample, the some college group did not score higher than the high school or less group. While this is somewhat unexpected, these mean differences are very small: The effect sizes of the differences between each of the three education level groups in the U.S. sample are negligible.

Table 2. Education Level Comparisons

Sample	Education level	*N*	Mean	*SD*	Difference	Effect size
French, France dialect	2 years college or less	163	-.03	.87	.31	.32
	3 years college or more	213	.28	1.06	–	–
English, U.S. dialect	High school or less	93	-.34	.91	.01	.01
	Some college	243	-.40	.94	.07	.07
	Bachelor's degree	307	-.33	1.02	–	–

Fairness and Group Comparisons

Fairness and equality of opportunity of assessment processes are important considerations for both legal and ethical reasons. In particular, the nature (e.g., magnitude and direction) of group differences in cognitive test performance that are consistently observed are relevant to test score interpretation. With regard to group score differences, administrators should be aware of the influence of possible test artifacts, relevance of the difference considering the testing purpose, and the potential for adverse impact.

First, group differences may be related to an artifact of the test rather than a true difference between groups. For example, content or question wording that is more salient for one group over another may create an unfair group performance difference.

In addition, a test should be relevant to the context of its use. When score differences reflect true group ability discrepancies, administrators should ensure that the tested ability is relevant to the job. For example, if a test was used to select people for a role where analogical reasoning is not relevant, it may result in unfair hiring decisions for examinees who did not perform well on the test. If a certain level of test scores shows evidence of relevance to job performance and business requirements, test score standards for selection should be set at that level.

Greater adverse impact is observed with higher cut scores than with lower cut scores. Group score differences and adverse impact are not the same, although they are closely related. Administrators should not assume a significant group difference always results in adverse impact or that a non-significant group difference guarantees there is no adverse impact. Adverse impact is situationally specific and is often related to more than the test alone. For instance, adverse impact can be influenced by the cut score used, the proportion of minority applicants, and the variance of group scores. It is the administrator's responsibility to monitor adverse impact on a regular, case-by-case basis. Adverse impact may be illegal unless it can be justified clearly in terms of the job requirements.

Sex Comparisons

Table 3 provides a comparison of the group mean scores of male and female examinees from the French, France dialect normative sample and the English, U.S. dialect sample. The results for the French, France dialect normative sample indicate that theta scores are slightly higher for males than for females, whereas the English, U.S. dialect sample shows essentially no difference between the two groups. The effect sizes are negligible for both differences, however.

The observed difference between the male and female group means in the U.S. sample is atypical of similar cognitive ability test study results. In most cases, the male group mean is higher than that of the female group (Kaufman et al., 2016; Raiford, Coalson, & Engi, 2014; Weiss, Saklofske, Coalson, & Raiford, 2010). However, the male group has slightly more high school graduates than the female group, which had more examinees with education levels greater than high school. Therefore, the male group mean may be an underestimate relative to what is typically obtained from larger samples.

Table 3. Sex Comparisons

Sample	Sex	*N*	Mean	*SD*	Difference	Effect size
French, France dialect	Female	197	.09	.93	.13	.13
	Male	181	.22	1.05	–	–
English, U.S. dialect	Female	319	-.34	.91	-.05	-.05
	Male	449	-.39	1.05	–	–

Race/Ethnicity Comparisons

Table 4 provides group mean score comparisons by race/ethnicity for the English, U.S. dialect sample. These comparisons are not available for the France, French dialect sample.

Each group mean is compared with that of the White group. The group means for examinees who are Black and for examinees who are Hispanic are slightly lower than the group mean for examinees who are White. The effect size of the difference between the White and the Black groups is small, and that of the White and Hispanic groups is negligible. The group mean for examinees who are Asian/Pacific Islander is higher than the group mean for examinees who are White; this difference produces a small effect size. The group mean difference between examinees who are White and examinees who described themselves as falling in a race/ethnicity category other than White, Black, Hispanic, or Asian/Pacific Islander is negligible.

Table 4. Race/Ethnicity Comparisons

Race/Ethnicity	*N*	Mean	*SD*	Difference	Effect size (compared to White non-Hispanic)
White (non-Hispanic)	579	-.37	1.01	–	–
Black, African American	67	-.61	1.01	.24	.24
Hispanic, Latino/a	42	-.49	.76	.12	.12
Asian/Pacific Islander	46	-.01	.94	-.36	-.36
Other	33	-.36	.88	-.01	-.01

Age Comparisons

Table 5 provides group mean score comparisons by age group for the French, France dialect sample, and the English, U.S. dialect sample. Each group mean is compared with that of the youngest group for that language. Performance on visual inductive reasoning tasks typically begins to decline sometime in the mid-30s (Wechsler, 2008).

For the French, France dialect sample, the highest score is in the group aged 25–29. The groups aged 25–29 and 30–34 obtained higher means than did the group aged 16–20, and these differences produced small effect sizes. The group aged 40–49 obtained the same mean as the youngest age group. The oldest group (aged 50–59) obtained a slightly lower mean than the group aged 16–20; however, the effect size was negligible. These results generally align with expectations.

For the English, U.S. dialect sample, the same pattern did not hold. Peak performance occurred in the group aged 40–49. The group aged 35–39 had the second highest mean. The groups of examinees younger than age 35 and those aged 60 and older produced noticeably lower means in relation to examinees aged 40–49. Comparisons of the group mean for those aged 16–24 to the group means for the two oldest age groups produced negligible effect sizes. Further investigation indicated that relative to the older age groups, the younger age groups had greater proportions of examinees with lower education levels. This age-related sample composition difference likely explains the unusual pattern of results.

Table 5. Age Comparisons

Sample	Age range in years	*N*	Mean	*SD*	Difference	Effect size (compared to youngest age band)
French, France dialect	16–20	39	.00	.92	–	–
	21–24	76	-.08	.99	.08	.08
	25–29	74	.40	.97	-.40	-.42
	30–34	67	.27	1.04	-.27	-.27
	35–39	53	.35	1.02	-.35	-.36
	40–49	47	.00	.96	.00	.00
	50–59	22	-.12	.81	.12	.14
English, U.S. dialect	16–24	69	-.41	1.05	–	–
	25–29	185	-.60	1.02	.19	.18
	30–34	199	-.42	.94	.01	.01
	35–39	122	-.15	.94	-.26	-.26
	40–49	107	-.10	.98	-.31	-.31
	50–59	55	-.28	.96	-.13	-.13
	60 and older	29	-.58	1.08	.17	.16

References

Aguinas, H., & Smith, M. A. (2007). Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. *Personnel Psychology, 60,* 165–199.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Americans With Disabilities Act of 1990, Titles I & V (Pub. L. 101–336). United States Code, Volume 42, Sections 12101–12213.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78*, 387–409.

Carroll, J. B. (1993). *Human cognitive abilities.* New York, NY: Cambridge University Press.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1): 155–159.

Colman, A. M. (2016). Race differences in IQ: Hans Eysenck's contribution to the debate in the light of subsequent research. *Personality & Individual Differences*, *103*,182–189. https://doi.org/10.1016/j.paid.2016.04.050

Corengia, A., Pita, M., Mesurado, B., & Centeno, A. (2013). Predicting academic performance and attrition in undergraduate students. *Liberabit*, *19*(1), 101–112.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York, NY: Harper & Row.

Deary, I. J., & Johnson, W. (2010). Intelligence and education: Causal perceptions drive analytic processes and therefore conclusions. *International Journal of Epidemiology, 39*(5), 1362–1369. https://doi.org/10.1093/ije/dyq072

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*(1), 13–21.

Diaz-Morales, J. F., & Escribano, C. (2013). Predicting school achievement: The role of inductive reasoning, sleep length and morningness-eveningness. *Personality and Individual Differences, 55*(2), 106–111.

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement, 27*(6), 440–458. https://doi.org/10.1177/0146621603258786

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions* (2nd ed.). New York, NY: Routledge.

Huepe, D., Roca, M., Salas, N., Canales-Johnson, A., Rivera-Rei, A. A., Zamorano, L.,...Ibañez, A. (2011). Fluid intelligence and psychosocial outcome: From logical problem solving to social adaptation. *PLoS ONE, 6(*9), e24858. https://doi.org/10.1371/journal.pone.0024858

Hunt, E., & Madhyastha, T. M. (2012). Cognitive demands of the workplace. *Journal of Neuroscience, Psychology, and Economics, 5*(1), 18–37. https://doi.org/10.1037/a0026177

Johnson, W., Deary, I. J., & Iacono, W. G. (2009). Genetic and environmental transactions underlying educational attainment. *Intelligence, 37*(5), 466–478. https://doi.org/10.1016/j.intell.2009.05.006

Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (2016). *Intelligent testing with the WISC–V.* Hoboken, NJ: Wiley.

Kaufman, S. B., Reynolds, M. R., Liu, X., Kaufman, A. S., & McGrew, K. S. (2012). Are cognitive *g* and academic *g* one and the same *g*? An exploration on the Woodcock-Johnson and Kaufman tests. *Intelligence, 40*, 123–138. https://doi.org/10.1016/j.intell.2012.01.009

Kuncel, N. A., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science, 19*(6), 339–345.

Kuncel, N. R., Ones, D. S., & Sackett, P. R. (2010). Individual differences as predictors of work, educational, and broad life outcomes. *Personality and Individual Differences, 49,* 331–336. https://doi.org/10.1016/j.Paid.2010.03.042

Lang, J. W. B., Kersting, M., Hulsheger, U. R., & Lang, J. (2010). General mental ability, narrower cognitive abilities, and job performance: The perspective of the nested-factors model of cognitive abilities, *Personnel Psychology, 63*, 595–640.

Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish Enlistment. *American Economic Journal: Applied Economics, 3*, 101–128.

Linn, M. C., & Peterson, A. C. (1985). Emergence and characterization of sex differences in spatial abilities: A meta-analysis. *Child Development, 56*, 1479–1498. https://doi.org/10.1111/1467-8624.ep7252392

Lipnicki, D. M., Crawford, J. D., Dutta, R., Thalamuthu, A., Kochan, N. A., Andrews, G., . . . Lam, L. W. (2017). Age-related cognitive decline and associations with sex, education and apolipoprotein E genotype across ethnocultural groups and geographic regions: A collaborative cohort study. *Plos Medicine, 14*(3), 1–21. https://doi.org/10.1371/journal.pmed.1002261

Maeda, Y., & Yoon, S. Y. (2013). A meta-analysis on gender differences in mental rotation ability measured by the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R). *Educational Psychology Review, 25*, 69–94. https://doi.org/10.1007/s10648-012-9215-x

Meade, A. W., & Fetzer, M. (2009). Test bias, differential prediction, and a revised approach for determining the suitability of a predictor in a selection context. *Organizational Research Methods, 12*, 738–761.

Meade, A. W., & Tonidandel, S. (2010). Not seeing clearly with Cleary: What test bias analyses do and do not tell us. *Industrial and Organizational Psychology, 3*, 192–205.

National Center for O*NET Development. *Abilities: Inductive Reasoning*. Retrieved June 5, 2018, from https://www.onetonline.org/find/descriptor/result/1.A.1.b.5?a=1

Nelson, J. M., Canivez, G. L., & Watkins, M. W. (2013). Structural and incremental validity of the Wechsler adult intelligence scale–fourth edition with a clinical sample. *Psychological Assessment, 25*(2), 618–630. https://doi.org/10.1037/a0032086

Pearson. (1991). *Differential Aptitude Tests for Personnel and Career Assessment Technical Manual.* NCS Pearson: Author.

Raiford, S. E., Coalson, D. L., & Engi, M. D. (2014). WPPSI–IV score differences across demographic groups. In S. E. Raiford & D. L. Coalson, *Essentials of WPPSI–IV assessment* (pp. 213–216). Hoboken, NJ: Wiley.

Raven, J. C. (1962). *Advanced progressive matrices*. London, England: Lewis.

Raven, J., Raven, J. C., & Court, J. H. (2000). *Manual for Raven's progressive matrices and vocabulary scales. Section 3: The standard progressive matrices*. San Antonio, TX: The Psychological Corporation.

Rushton, J. P., & Ankney, C. D. (2009). Whole brain size and general mental ability: A review. *International Journal Of Neuroscience*, *119*(5), 692–732. https://doi.org/10.1080/00207450802325843

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). San Diego, CA: Author.

Schmidt, F. L. (2014). A general theoretical integrative model of individual differences in interests, abilities, personality traits, and academic and occupational achievement: A commentary on four recent articles. *Perspectives on Psychological Science, 9*(2), 211–218. https://doi.org/10.1177/1745691613518074

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*(1), 162–173. https://doi.org/10.1037/0022-3514.86.1.162

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–163). New York, NY: Guilford.

Smith, P., & Whetton, C. (2011). *Tests d'aptitudes generales*. Montreuil, France: Pearson France - ECPA.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16* (1), 1–9.

Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

U.S. Department of Labor. (1999). *Testing and assessment: An employer's guide to good practices*. Washington, DC: Author.

U.S. Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 3829385–309.

Voyer, D., Voyer, S. D., & Saint-Aubin, J. (2017). Sex differences in visual-spatial working memory: A meta-analysis. *Psychonomic Bulletin & Review, 24,* 307–334. https://doi.org/10.3758/s13423-016-1085-7

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22–29.

Wechsler, D. (2008). *Wechsler adult intelligence scale* (4th ed.). Bloomington, MN: Pearson.

Weiss, L. G., Saklofske, D. H., Coalson, D., & Raiford, S. E. (Eds.) (2010). *WAIS–IV clinical use and interpretation*. San Diego, CA: Academic Press.

Appendix A. Test Log

The following page may be photocopied for use in your organization.

Test Log

Organization	
Purpose of Testing	Selection / Development / Appraisal
Test(s) Used (Circle all that apply.)	Numerical Sequences Abstract Reasoning ANAS ANRA NDIT Numerical Calculation Verbal Reasoning Space Relations Watson-Glaser Ravens Adaptive Other:
Test Administrator(s)	
Date	
Start time	
Finish time	

Candidate List

	Candidate's Name	Test	Retest		Candidate's Name	Test	Retest
1.				6.			
2.				7.			
3.				8.			
4.				9.			
5.				10.			

Disturbances/Unusual Occurrences

.