PXT SELECT™
A Wiley Brand

PXT Select™ Research Report

Version 1.0

WILEY

Contents

List of Tables

Overview of this Research Report

The purpose of this report is to provide research study results for the *PXT Select™* assessment. The first section includes background and research on the assessment, specifically on the job fit model. The next sections cover the assessment scale measures and scoring, the assessment validation process, and reliability and validity results. The final sections provide information about the development of the Performance Model library and research supporting compliance with legal issues surrounding fairness and adverse impact.

The Job Fit Model

PXT Select evaluates a person's cognitive abilities, behavioral traits, and interests, reflecting the three areas of the person-environment fit model. The decision to focus on these three areas for *PXT Select* is based on findings in the field of occupational research, dating back to 1909 with Parsons' Tripartite model for vocational direction (Parsons, 1909). The model proposes three areas for choosing a vocation, including an accurate understanding of one's aptitudes, personal abilities, and interests; knowledge of the job market; and an objective view of the relationship between one's own traits and available jobs.

Parsons' initial framework for decision making in vocational counseling evolved into Williamson's 1939 job-matching concept called the Trait and Factor Theory of Occupational Choice (Williamson, 1939). This model gives the job match process increased precision by applying statistical measurement to determine the degree of relationship between individuals' traits and the factors required for success in the work environment. Ackerman and Heggestad's work in the late 1990s shows that intelligence, personality, and interests are part of trait complexes that help explain the total person, and that these areas help inform successful occupational choice (Ackerman & Heggestad, 1997). These foundational ideas spurred continued research into occupational success, leading to contemporary person-environment fit theories.

Person-environment fit models suggest that people are drawn to environments that allow them to express their traits behaviorally, that outcomes affected by person-environment fit are significant and positively related to higher degrees of fit, and that person-environment fit can be adapted over time (Rounds & Tracey, 1990). Early clinicians, counselors, and occupational researchers, and since then industrial/organizational psychologists, career coaches, and HR practitioners, all agree that individuals

perform best when they hold positions best suited to their traits, that is, when the person-environment fit is high. See Figure 1 for a schematic representing person-environment fit models.

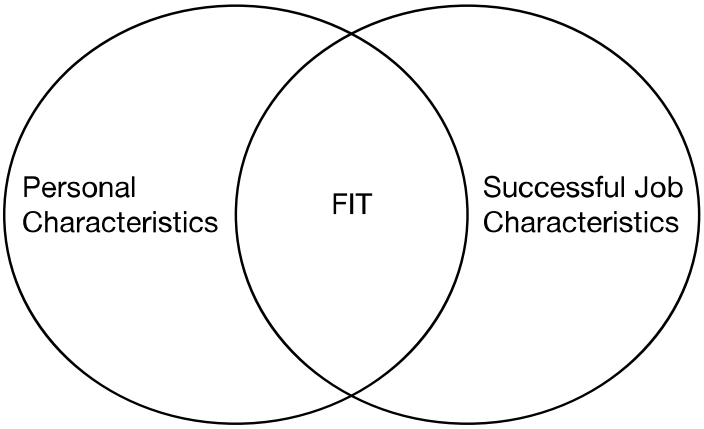

Figure 1. Person-Environment Fit Model

Taken together, these models provide a comprehensive conceptual framework for an understanding of how matching someone's personality and ability with job requirements can lead to positive performance outcomes. They reinforce the need for objectivity in determining how well someone's traits match the requirements at work, and hence, the degree of fit someone has with a job. *PXT Select™* objectively measures participants' cognitive, behavioral, and interest traits and describes the degree to which these traits fit the position, which is represented by a tool called the Performance Model. *PXT Select* brings the foundational theories of job performance to today's organizations, supported by over a century of research in workplace success.

Assessment and Scoring

Thinking Style

The cognitive section of *PXT Select,* known as the Thinking Style section, asks participants to answer multiple-choice questions in four areas of knowledge and reasoning. These responses are used to calculate scores on the following five scales that reflect verbal and numerical skill and reasoning:

Composite Score is a reflection of overall learning, reasoning, and problem-solving potential, a combined and standardized score of all four cognitive scales.

Verbal Skill is a measure of vocabulary, for which respondents identify synonyms to words.

Verbal Reasoning measures the ability to use words in reasoning and problem solving, for which respondents complete analogies.

Numerical Ability measures numeric calculation abilities, for which respondents use numeric formulas to perform calculations.

Numeric Reasoning measures the ability to use numbers as a basis in reasoning and problem solving, for which respondents perform calculations to solve numerical problems.

Behavioral Traits

The Behavioral Traits section of *PXT Select™* asks participants to agree or disagree with statements of work-related preferences and behaviors in nine distinct areas. These responses are used to determine the way in which the individual expresses each trait. The nine scales are as follows:

Pace measures a preference for the overall rate of task completion with items referring to urgency and liveliness or steadiness.

Assertiveness measures the preference for expressing opinions and the need for control with items referring to persuasiveness and taking charge or being more unassuming.

Sociability measures the desire for interaction with others, with items asking respondents about being outgoing and meeting new people or being more reserved and quiet.

Conformity measures respondents' attitude on policies and supervision, with items about challenging things and showing discontent or following rules and policies.

Outlook measures the type of anticipation of outcomes and motives respondents have, whether more skeptical and practical or trusting and optimistic.

Decisiveness measures the preference for speed or caution to make decisions, with items asking respondents how much time they take to make decisions.

Accommodation measures the inclination to tend to others' needs and consider their ideas, asking respondents about expressing opinions and taking a stand or promoting harmony and agreement.

Independence measures respondents' preference for instruction and guidance, with items referencing procedures, preference for flexibility, and attitude toward status check-ins.

Judgment measures respondents' basis for forming opinions and making judgments, asking them the role of instincts or logic in their decision making.

Adaptive Testing Format

The adaptive testing format is used for the cognitive and behavioral sections of the assessment. Adaptive testing maximizes the precision in the measurement of respondents' true score of the cognitive or behavioral measures, while minimizing the number of questions required. Over the course of the assessment, increasingly more difficult cognitive questions are given in response to correct answers, and easier questions are given following incorrect answers. Behavioral items are presented in a similar format, with increasingly strongly worded statements given with each endorsed item. Item endorsement occurs when respondents indicate they do, in fact, feel that they express the behavioral trait in one way or another. Questions are presented in one direction of difficulty, or behavioral expression, or the other until a stable result is reached. This result represents the best measurement of the individual's true nature on that scale. Results on the nine Behavioral Traits are reported on a continuum scale between two endpoints that describe opposing expressions of each scale measure. Results on the five Thinking Style scales are reported on a 10-point scale, with higher values indicating a higher level of knowledge and skill in each area.

Interests

The Interests section of *PXT Select™* asks participants to rate their level of enjoyment of various activities, to measure their degree of interest in six different areas. These areas are based on John Holland's RIASEC classification for occupational interests: Realistic, Investigative, Artistic, Social, Enterprising, and Conventional (Holland, 1959). Responses are used to determine respondents' relative degree of enjoyment among the six interests. The report shows the individual's interests ranked from

highest to lowest preference. In the case where multiple interests are similarly enjoyed, a tie between the interests is shown. The six interests are as follows:

Enterprising, also labeled Enterprising in Holland's model, suggests the enjoyment of leadership, presenting ideas, and persuading others.

Financial/Administrative, or Holland's Conventional interest, suggests enjoyment working with numbers and organizing information.

People Service, or Holland's Social interest, suggests enjoyment of collaboration, compromise, and helping others.

Technical, or Holland's Investigative interest, suggests enjoyment of learning technical material, interpreting complex information, and solving problems.

Mechanical, or Holland's Realistic interest, suggests enjoyment of building and repairing things, working with the hands using machinery and tools.

Creative, or Holland's Artistic interest, suggests enjoyment of imaginative and artistic activities.

Paired Comparison Format

The paired comparison format is used for the Interests section of the assessment. In the paired comparison format, respondents compare their level of enjoyment in a pair of activities. The activities within each pairing represent two distinct interest scales for *PXT Select™.* Respondents indicate whether they would highly prefer one activity or the other, or whether they would somewhat prefer one activity or the other. Over the course of the Interests section, successive preferences for one interest activity or the other are compiled and scored. The more activities a respondent indicates he or she would enjoy for a given interest scale, the higher that interest will appear in a rank order of interests on the reports. Any ties between interests are also reported. Report feedback is based on the position of the interest scale in the ordered list. Some types of feedback also reflect any matches between the respondent's top three interests and the interests reflected in the Performance Model.

Distortion

Because *PXT Select™* is a self-report assessment battery, it may be difficult for end users of the assessment information to tell if the participant was candid while responding to Behavioral Traits items. This can be true of any assessment in which respondents are asked to self-report, but it is a particular concern for any high-stakes assessment such as those used in workforce decisions. In hiring or other professional contexts, and surely during any hiring process evaluation exercise, the respondent may be tempted to give an impression he or she feels would look favorable. In this case, the individual may respond to behavioral items in such a way that assessment results may not reflect his or her actual approach to workplace situations. To detect the chance of this having occurred, the *PXT Select* Distortion measure provides a method for determining the respondent's level of candor in response to the assessment's behavioral items.

However, it is crucial to ensure that the Distortion measure does not unduly identify those who may actually be candid in their responding, but may hold uncommon perspectives. To prevent this from happening, the Distortion measure is validated, with the goal of minimizing the chance of incorrectly reporting Distortion for someone who was candid in his or her responses.

Distortion Detection

To detect the possibility of Distortion, a number of items are included in the behavioral section that ask respondents for their perspectives on various situations. These situations are designed to be indistinguishable to the respondent from items that measure other behavioral constructs. However, these items reference situations that, if the respondent is candid, may not give the absolute best impression of him or her. Even so, the items were chosen for measuring Distortion because the vast majority of the U.S. population would be expected to answer them in the same way. Said another way, the expected answer choice for each Distortion item is the consistent response expected for the vast majority of others in the population. (See Table 1 for the item response rates for each Distortion item.) For any one individual, the more Distortion items he or she answers in the expected, consistent way, the more likely his or her responses to other behavioral items were made with a similar degree of candor. The Distortion measure is not presented in the adaptive test format. All Distortion items are asked of every respondent.

Table 1. Distribution of Distortion Item Response Rates

Item*	Response	Frequency	Percent
1	Non-Loading**	1,966	91.2
	Loading**	190	8.8
2	Non-Loading	1,953	90.6
	Loading	203	9.4
3	Non-Loading	2,084	96.7
	Loading	72	3.3
4	Non-Loading	2,041	94.7
	Loading	115	5.3
5	Non-Loading	2,025	93.9
	Loading	131	6.1
6	Non-Loading	1,508	69.9
	Loading	648	30.1
7	Non-Loading	1,812	84.0
	Loading	344	16.0
8	Non-Loading	2,063	95.7
	Loading	93	4.3
9	Non-Loading	1,765	81.9
	Loading	391	18.1

*Actual item numbers in the assessment are not revealed.
**Loading responses are those that are not considered representative of an open level of disclosure; non-loading responses are those that represent a more candid level of disclosure.

Methodology and Results

The validation of the Distortion measure ensures that it will only be reported in cases where there is an acceptable degree of confidence that the behavioral item responses may not be candid. To determine the validity of the Distortion measure, the distribution of Distortion measure results of a test sample group of 2,156 individuals was analyzed. This was to make sure that a very limited number of respondents would be associated with a Distortion indicator, with a similar measure of confidence in the accuracy of the result for them.

Table 2 shows the number of individuals in the test group of 2,156 respondents who answered in the consistent way, for between zero and nine items. Just over 50% of the total sample responded to all

Distortion items in an expected, consistent way. Slightly more than 25% of the sample responded to only one Distortion item in an unexpected way. Finally, just less than 1% of the sample responded inconsistently to almost all Distortion items.

Table 2. Distribution of Distortion Measure Results

No. of Consistent Responses	Frequency	Percent
0	9	0.4
1	10	0.5
2	9	0.4
3	28	1.3
4	43	2.0
5	65	3.0
6	95	4.4
7	242	11.2
8	551	25.4
9	1,104	50.9
Total	2,156	

In light of these results, a lack of candor in responding to behavioral items is likely present for only a fraction of a percent of those who complete the *PXT Select™* assessment.

Scoring

Assessment Scale Scoring

In the *PXT Select* scoring system, assessment scores are converted to scaled scores and reported as Standard Ten (STEN) scores on the Thinking Style section. STEN scores are used to define the relative level of knowledge and skill on a Thinking Style scale for a given respondent. The STEN scale is a 10-point scale ranging in value from 1 to 10 and is normally distributed across the U.S. working population. Roughly two-thirds (68%) of the scores will fall between 4 and 7, which is within one standard deviation of the STEN scale mean score. STEN scores are used to define the relative level of a given *PXT Select* trait found for the respondent. Higher STEN scores on Thinking Style scales may be appropriately interpreted as higher degrees of skill and knowledge in the areas covered by the scales.

The process of converting assessment scores to STEN scores involves the use of a STEN score conversion table. This table is created during the development of the assessment through a process of norming. Norming scores involves slicing the scores of a sample group into 10 distinct groupings. Each score grouping represents a segment of the population scoring higher or lower than other segments. The frequency of scores occurring for each grouping follow a normal distribution, more commonly known as a bell curve, of scores. STEN scores representing the 10 groupings of scores reflect values from 1 to 10 on the scale.

The norm sample is the sample group whose *PXT Select™* scores are used to create the STEN score conversion table. They represent the larger population of respondents who will complete *PXT Select* in some workforce decision process. Because the norm sample represents the larger U.S. working population, a large variety of jobs and industries are represented by those in the norm group. These range from unskilled labor to highly specific professional and technical jobs, covering every job family in the U.S. Department of Labor's O*Net database. The following tables show the distribution of groups within the norm sample for gender, age, ethnicity, and educational levels (see Tables 3–6). The overall sample group was comprised of 1,130 individuals, some of whom answered some of the demographic questions, some of whom answered others. This accounts for the differences in totals in the following tables.

Table 3. Distribution of Norm Sample Gender Groups

	Frequency	**Percent**
Female	555	50.6
Male	541	49.4
Total	1,096	100.0

Table 4. Distribution of Norm Sample Age Groups

	Frequency	Percent
0–17	4	.4
18–25	226	20.2
26–30	144	12.8
31–35	128	11.4
36–40	124	11.1
41–45	116	10.3
46–50	126	11.2
51–55	114	10.2
56–60	75	6.7
61–65	43	3.8
66+	21	1.9
Total	1,121	100.0

Table 5. Distribution of Norm Sample Ethnic Groups

	Frequency	Percent
White, not of Hispanic origin	817	72.9
Asian or Pacific Islander	55	4.9
Black, not of Hispanic origin	146	13.0
Hispanic	73	6.5
American Indian or Alaskan Native	5	.4
Other	25	2.2
Total	1,121	100.0

Table 6. Distribution of Norm Sample Education Levels

	Frequency	Percent
Some High School	22	2.0
High School Graduate	146	13.0
Technical or Trade School	55	4.9
Some College	273	24.2
College Graduate	455	40.4
Graduate or Professional Degree	175	15.5
Total	1,126	100.0

The norm sample represented the demographic characteristics of gender, age, and ethnicity in proportions that approximate the U.S. Census Bureau's records for 2010 and 2015. Norm sample participants also represented various education levels. Not surprisingly for respondents of a workforce assessment such as *PXT Select™,* there are slightly higher proportions of high school and college graduates in this sample of the U.S. working population than those given by the U.S. Census Bureau for the overall, general population. Additionally, because *PXT Select* is intended for use in any workforce decision context, including selection and professional development, the norm sample is comprised of equal proportions of applicants who completed the assessment as part of an organization's hiring process and incumbent employees asked by their organization to complete the assessment.

Percent Fit Calculation

Percent fit is a numerical result generated by comparing a respondent's scores for each of the Thinking Style, Behavioral Traits, and Interests sections to the ideal scores for the Performance Model. Separate percent fit values are provided for each section, which all contribute to the Overall Fit result. The more similar the respondent's results are to the result ranges indicated by the Performance Model, the higher the values will be for Overall Fit, Thinking Style Fit, and Behavioral Traits Fit. Similarly, the more closely the respondent's interests resemble those in the Performance Model, the higher the values will be for Overall Fit and Interests Fit. The Overall and sectional percent fit values range from 25% to 95%. Although the assessment calculates ranges from 1% to 100%, values below 25% and above 95% tend to be too extreme to be useful and don't really help predict greater or less success on the job. The Overall Fit result is especially useful to identify individuals who may be a good fit for the job. The percent fit for the three subsections can be used for a more in-depth analysis of job fit. The higher the reported percent fit, the higher the expectancy the individual will fit well into the job under

consideration. The Overall Fit value is the key metric that end users should reference for workforce decisions, for the portion of the decision that is influenced by the assessment results.

Overview of the Validation Process

Occupational assessments are used to measure abilities and traits that are relevant for most workplaces, such as verbal and numerical skills, tendency to follow rules and procedures, and sense of urgency when completing tasks. This type of information can be invaluable for organizations, and the quality of the information hinges on the accuracy, validity, and reliability of the assessment measures. So then, how does one know if the assessment, and by extension the information about the respondent, is, in fact, accurate, valid, and reliable? There are certain criteria that are used in determining whether the assessment is valid, as well as guidelines and standards for knowing how these criteria are satisfied. The psychological measurement community has established a set of statistical practices and values the assessment should meet to be considered valid for precise and stable measurement and scoring (American Educational Research Association, et al., 1999). In addition, the U.S. Department of Labor (DOL) provides guidelines for employers in evaluating reliability and validity of assessments, making sure that they are appropriate for use in workforce decisions (U.S. Department of Labor, 2000).

In 2000, the U.S. Department of Labor published *Testing and Assessment: An Employer's Guide to Good Practices*. This guide gives employers a complete view of all the considerations they need to take into account when including assessments as part of their workforce planning. It includes what the DOL calls "Principles of Assessment," which are 13 of the most important factors for any organization planning to use personnel assessments. Many organizations may not have deep knowledge or expertise in human resource planning, particularly in the field of assessments. This guide provides organizations the information they need to make effective and legally responsible decisions, and for making sure that the assessment and the associated reports they use are scientifically tested and fair for all respondents.

The U.S. Department of Labor states that a good test (1) measures traits or skills consistently or reliably, and (2) does indeed measure what it claims to measure. In the field of psychometrics, these considerations are known as **reliability** and **validity**, respectively.

Reliability refers to an assessment's measurement of various abilities, traits, and interests in a consistent and dependable way, both for different individuals and over different periods of time. In

short, reliability tells us that everyone is being measured the same way, and that if individuals' results differ, it is due, for instance, to behavioral differences between them, not differences in the assessment. **Validity** refers to an assessment's ability to measure the scales accurately and in the way it purports to measure them. It tells us, in a number of ways, how strongly the assessment items measure what the scale name, definition, conceptual endpoints, and report feedback describe.

It is crucial to know, however, that while there are standards and guidelines around appropriate levels of reliability and validity, when evaluating reliability and validity, there is not a specific, required type of test, nor a pass/fail value above or below which a test must adhere. Rather, developing an assessment includes an extensive process where researchers examine a sample group's scores. This sample group represents the larger population that will be taking the assessment. The studies result in numerous lines of evidence that either support or refute the conclusion that the assessment, and the resulting information about respondents, is valid and reliable for use as intended.

The goal of the following sections is to present the numerous lines of evidence for the reliability and validity of the *PXT Select™* assessment in measuring 20 cognitive, behavioral, and interest areas. This section will demonstrate that *PXT Select* was developed with all of the U.S. Department of Labor's Principles of Assessment for test quality in mind, and that the results meet the statistical guidelines they offer for acceptable test quality.

Reliability

The DOL Principle of Assessment related to reliability states that employers should "use only *reliable* assessment instruments and procedures. In other words, use only assessment tools that provide dependable and consistent information" (U.S. Department of Labor, 2000, p. 3-3). Two types of reliability tests were conducted for *PXT Select* to provide evidence that the assessment is dependable and provides consistent information—test-retest and internal consistency. Sample groups of approximately 300 to 600 individuals each participated in the reliability studies for the development of *PXT Select.*

Sample Group Characteristics

The participant sample represents the appropriate proportions of age, gender, ethnicity, and education levels found in the overall U.S. population, based on 2010 and 2015 U.S. Census Bureau statistics. They were employed in 30 different job classifications, covering 20 different industries. Participants

completed the assessments for the reliability studies online, in an unproctored setting, identical to the conditions under which the *PXT Select™* assessment is administered.

Test-Retest Reliability

Test-retest reliability tells us the stability of measurement results over the passage of time, as well as the stability of the scale measure, even given all the factors that can potentially affect results. To determine this type of reliability, researchers look at assessment results across repeated administrations of the assessment to the same individual, by asking a sample group of individuals to complete the entire assessment twice. They complete the initial assessment, and their results are determined by the scoring system. Then, after a period of time, the group completes the same assessment a second time, and their results are determined by the system once again. Researchers recognize that if an individual's results from his or her first assessment are similar to the results on the second assessment, then the assessment measures the constructs in the same way over time. In other words, the more consistent individuals' scores are between the two assessments, the more stable the assessment is in measuring the scales over time.

Participant responses to certain items, of course, may change over time, to varying degrees among different people. There are a number of reasons this may occur. In the case of a behavioral item, the respondent might be torn between two responses, and on one day responds that he or she agrees with the item, but on the second administration, feels equally inclined to disagree with the item. Over a period of even weeks, moods change, circumstances of everyday life change, knowledge and perspectives may change—all factors that can affect responses on a self-report behavioral measure. In addition, cognitive tests are always susceptible to changes in results over time, again, even over a period of weeks, if respondents learn new verbal or numerical concepts, look up answers to remembered items from the first assessment, or become subject to practice or fatigue effects. Given all these types of risks to result stability, the assessment and resulting scoring scheme must be developed in such a way as to minimize each of these possible effects.

Test-retest reliability, and its calculations for measurement error, is a common way to demonstrate that the assessment does, indeed, minimize the effect of these myriad factors on assessment results. This is especially important for adaptive tests, where changes in responding can change the item set presented to a particular individual for a given scale. Adaptive tests should measure the same construct reliably, even if the item set that a particular respondent receives for a given scale changes. In this way, with acceptable values of test-retest reliability, the adaptive assessment can be considered a stable

measure of the scales it contains. Therefore, the test-retest analyses were conducted for the *PXT Select™* adaptive format sections, Thinking Style and Behavioral Traits.

Test-retest reliability is reported as a correlation between results from the first and second assessment for each respondent. This value can vary between 0.0 and 1.0, and the closer to 1.0 the correlation is, the more strongly related the results from each test administration are. Guidelines offered by the DOL to interpret the quality of a test based on reliability estimates, which also follow the scientific community's accepted practices, suggest that coefficients above .70 are acceptable. The average test-retest reliabilities for the *PXT Select* Thinking Style and Behavioral Traits sections exceed .70, as shown in Tables 7 and 8.

In addition, the U.S. Department of Labor suggests that employers should also consider the standard error of measurement (SEM) of reliability statistics when considering assessments to help make personnel decisions. The SEM "gives the margin of error that you should expect in an individual test score because of imperfect reliability of the test. The SEM represents the degree of confidence that a person's 'true' score lies within a particular range of scores" (U.S. Department of Labor, 2000, p. 3-5). The DOL guidelines further note that the SEM is a measure of the accuracy of test result reliability, where the smaller the measurement error, the more accurate the measurement is. That is, the smaller the error for a reliability statistic, the better the chance that the reliability statistic is stable. Also, the better the chance that the reliability value is a good indicator of consistency in results for many administrations of the test among many different individuals or groups of individuals. In sum, the SEM tells how stable and accurate the reliability statistic is for each assessment scale. Therefore, the SEM was calculated for each reliability statistic, both for test-retest and internal consistency reliability values.

Methodology

Test-retest reliability studies were conducted separately for the Thinking Style and Behavioral Traits sections of *PXT Select.* That is, one group completed the Thinking Style section twice, while a different group completed the Behavioral Traits section twice. Because there is no result interpretation in the *PXT Select* reports that relies on any interaction among any cognitive and behavioral scales, this partitioning is appropriate. The median time between the first and second test administrations was 25 days, with a minimum time between tests of 12 days and a maximum of 41 days.

Thinking Style

Test-retest correlations for the results on the four cognitive scale measures were calculated for a sample of 324 respondents. In addition, the SEM was calculated, as suggested by the DOL, for employers to evaluate the degree of error associated with the reliability statistic. The scales on the *PXT Select™* Thinking Style section demonstrate good consistency, as shown by the values listed in Table 7, with an average reliability coefficient of .77 and an average SEM of 1.063.

Table 7. Test-Retest Correlations for PXT Select Thinking Style Scales

Scale	Reliability	SEM
Verbal Skill	0.72	1.176
Verbal Reasoning	0.75	0.966
Numerical Ability	0.78	1.139
Numeric Reasoning	0.82	0.973
All correlations significant at the .001 level (2-tailed).		

Behavioral Traits

Test-retest correlations and their associated SEM for the results on the nine behavioral scale measures were calculated for a sample of 300 respondents. The scales on the *PXT Select* Behavioral Traits section demonstrate good consistency, as shown by the values listed in Table 8, with an average reliability coefficient of .72 and an average SEM of 1.002.

Table 8. Test-Retest Correlations for PXT Select Behavioral Traits Scales

Scale	Reliability	SEM
Pace	.84	0.684
Assertiveness	.82	0.954
Sociability	.70	1.011
Conformity	.73	1.221
Outlook	.58	0.922
Decisiveness	.63	1.089
Accommodation	.74	1.106
Independence	.73	0.998
Judgment	.70	1.038
All correlations significant at the .001 level (2-tailed).		

Internal Consistency

Internal consistency evaluates another aspect of assessment reliability, the degree of correlation among questions that profess to measure the same thing. That is, each of the cognitive and behavioral scales of *PXT Select™* is measured using a set of different items referencing some aspect of the cognitive or behavioral measure. This is especially important for behavioral scales, where different statements are intended to measure the same scale, for instance, Pace. Statements to which respondents are asked to agree or disagree may vary in context, such as asking whether they tend to be calm, like to be in the center of the action, or never take a break. Researchers recognize that people tend to respond to these items in a predictable, consistent way. For example, if people do report that they tend to be calm, then they also probably work methodically. Both of these aspects of the pace at which someone works are reflected in the items on a single scale. If many people in the test group respond similarly consistently across the items, we know that the items are consistent with each other. That is, that responses to the items correlate with each other to some degree. Taken as a set, where individuals respond to items in a predictable, consistent way, this means the scale measure itself is a consistent, reliable measure of the construct. A statistic called Cronbach's alpha (Cronbach, 1984) is usually regarded as the best method of indicating the degree of consistency among items of a scale, or the internal consistency reliability.

Cronbach's alpha is a degree of correlation among the items of a scale. This value typically falls between 0.0 and 1.0, and the closer to 1.0 the value is, the more strongly related the items are to each other. As a special use of the analysis to evaluate the degree to which all of the items in the pool are measuring a single construct, Cronbach's alpha is computed for the entire item bank available for each scale. The same guidelines offered by the DOL in interpreting reliability coefficients apply for Cronbach's alpha as well, suggesting coefficients exceeding .70 as acceptable, above .80 as good, and above .90 as excellent. The average alpha coefficients for internal consistency for the *PXT Select* Thinking Style and Behavioral Traits sections exceed .80, as shown in Tables 9 and 10.

Thinking Style

A test group of 324 respondents completed all item bank questions available for each Thinking Style scale. Alpha coefficients for the four cognitive scale measures were calculated for the entire item bank of each Thinking Style scale. The scales on the *PXT Select* Thinking Style section demonstrate good consistency, as shown by the values listed in Table 9, with an average internal consistency coefficient of .86.

Table 9. Internal Consistency of the PXT Select Thinking Style Scales

Scale	Alpha	No. of Items	SEM
Verbal Skill	.88	39	2.419
Verbal Reasoning	.84	45	2.708
Numerical Ability	.86	28	2.184
Numeric Reasoning	.87	38	2.470

Behavioral Traits

A test group of 300 respondents completed all item bank questions available for each Behavioral Traits scale. Alpha coefficients for the nine behavioral scale measures were calculated for the entire item bank of each Behavioral Traits scale. The scales on the *PXT Select™* Behavioral Traits section demonstrate good consistency, as shown by the values listed in Table 10, with an average internal consistency coefficient of .89.

Table 10. Internal Consistency of the PXT Select Behavioral Traits Scales

Scale	Alpha	No. of Items	SEM
Pace	.86	30	2.047
Assertiveness	.96	63	3.238
Sociability	.95	63	3.079
Conformity	.92	54	3.094
Outlook	.87	45	2.636
Decisiveness	.82	38	2.315
Accommodation	.90	50	2.949
Independence	.86	35	2.240
Judgment	.86	39	2.576

Interests

A test group of 608 respondents completed the Interests section of *PXT Select.* Alpha coefficients for the six interest scale measures were calculated for the entire item set for each interest scale. The scales on the *PXT Select* Interests section demonstrate good consistency, as shown by the values listed in Table 11, with an average internal consistency coefficient of .77.

Table 11. Internal Consistency of the PXT Select Interest Scales

Scale	Alpha	No. of Items	SEM
Enterprising	.81	14	4.037
Financial/Administrative	.76	15	4.093
People Service	.73	13	3.958
Technical	.76	14	3.543
Mechanical	.79	15	3.712
Creative	.77	15	3.965

Validity

There are many ways to describe the process of determining how closely an assessment measures what it purports to measure. People may ask if the assessment is appropriate for measuring what they need it to measure for their personnel decisions. Common ways to ask this are, “Has the assessment been validated?” “Has the validity been tested?” or “Is the assessment valid?” While this phrasing can serve, conversationally, as an overarching description for the entire process of carrying out the various studies assessment publishers conduct, the scientific meaning of the terms *valid*, *validation*, and *validity* is very specific.

Validity is a long-established and well-accepted principle of any scientific study, telling us how well the scales of an assessment measure what they say they do. For *PXT Select™,* the cognitive, behavioral, and interest scales are measuring unseen, abstract concepts about the level of people’s knowledge, their perspectives on work-related situations, and things that they would enjoy doing. These concepts can also be called constructs, which can be measured with specific examples of how the constructs could be expressed. These examples of the constructs are the assessment items.

We use the items to operationalize the constructs—that is, describing in the items how people would act in their environment if they, for instance, know how to calculate difficult problems, feel the need to be urgent, or enjoy working with their hands. If people respond to these examples in ways that are consistent with how we expect them to, we say that the items are operationalizing the constructs accurately. If the items are operationalizing the constructs accurately, one may conclude that the evidence does support the validation of the scales. In this way, the assessment validity has been tested and confirmed. Only after a very specific, scientific process to test the validity and draw these

conclusions can we address the questions, “Has the assessment been validated?” “Has the validity been tested?” or “Is the assessment valid?”

Those questions are not answered most accurately with a yes or no response. Rather, we present lines of evidence confirming validity until we are satisfied that validity has been tested as thoroughly as possible, and with strong enough validity values that we are satisfied that the assessment is valid for its intended purpose. Researchers have identified many different types of validity and many ways to test each different type. No one type of validity is better than another. Rather, we choose the types of validity that are most appropriate for our assessment, its intended usage, and its expected participants. The U.S. Department of Labor’s recommended practices for employers in determining test validation include two Principles of Assessment related to test validity.

The first principle related to test validity recommends that employers should “use only assessment procedures and instruments that have been demonstrated to be valid for the specific purpose for which they are being used” (U.S. Department of Labor, 2000, p. 3-6). That is, the groups that were used to conduct the validity studies should represent the target population for which the assessment is to be used. To satisfy this principle with *PXT Select™,* we identified a test group representing a cross-section of the U.S. population that would simulate a population of persons taking a personnel assessment as candidates for employment. The sample group characteristics approximate the demographic distribution of the U.S. population as measured by the U.S. Census Bureau in 2010 and in 2015, with respect to age, gender, ethnicity, and education level. In addition, a wide range of jobs and industries are represented that span the entire range of jobs and industries in which organizations may use *PXT Select.*

The second DOL Principle of Assessment for test validity suggests that employers should “use assessment tools that are appropriate for the target population” (U.S. Department of Labor, 2000, p. 3-6). The target population would be the intended respondents of *PXT Select,* namely, the segment of the U.S. population that would participate in a personnel assessment. The guidelines point out that there are a number of ways that the test group and the target population can be similar, and that for an assessment to have validity, they do not need to match on every aspect possible. Important considerations may include the types of occupations reported on by the assessment, cultural differences, and language barriers. The wide range of jobs and industries held by those in the test groups for *PXT Select* represent each one of the job categories and industries in the U.S. Department of Labor’s O*Net database of occupational information. Any cultural differences were mitigated by limiting

the sample to U.S. residents, and any language barriers were eliminated by requiring everyone to be native English speakers.

For the validity analyses, participants completed the assessments online, in an unproctored setting, identical to the conditions under which the *PXT Select™* assessment is administered. In sum, *PXT Select* was tested for validity on a sample of individuals highly representative of the intended participants who will complete the assessment.

If *PXT Select* scales that are expected to be related, such as Pace and Decisiveness, indeed are, then the test group's responses on both of these scales should be highly correlated. That is, as their Pace increases, so should their Decisiveness, because we think that someone's sense of urgency will be positively related to how quickly he or she makes decisions. This is called **convergent validity**, where two scales that are expected to converge, indeed, do. The correlation between the two scales, then, becomes the coefficient of convergent validity, validating our expectation of the relationship with a high validity coefficient test result. On the other hand, sometimes we expect two scales to be unrelated, where we do not hypothetically expect a strong relationship to exist. We expect this validity coefficient to be small, because we don't expect a strong relationship between the two. If this validity result is, in fact, small, then we have demonstrated criterion-related **discriminant validity**. Our criterion was the expectation that the scales are or are not related, and the convergent and discriminant validity results either confirm or dispute this expectation. If either convergent or discriminant validity exists, researchers present it as a line of evidence supporting scale construct validation.

Scale Intercorrelations

Convergent and discriminant validity are shown in scale intercorrelation tables for the Thinking Style, Behavioral Traits, and Interests sections. Convergent validity is supported when our expectations that the scales we think should be related are met. The correlations between these scales, as shown in the scale intercorrelation tables, should follow those expectations. If they do, convergent validity for the constructs we expect to be related is supported by the correlation values.

The correlations vary between -1.0 and 1.0. Positive values between 0.0 and 1.0 mean that both scales' results increase together; that is, the higher one scale's results are, the higher the other scale's results are. Negative values between 0.0 and -1.0 mean that as one scale's results increase, the other decreases. That is, the higher one scale's results are, the lower the other scale's results will be. Correlations near -1.0 and 1.0 mean the relationship is strong, whereas correlations near 0.0 mean the

relationship is weak to non-existent. Just as for reliability, the DOL has recommended guidelines, which also follow the scientific community's accepted practices, for interpreting the quality of a test based on validity estimates. These guidelines suggest that, for a study sample size of 32, coefficients above ± .11 are acceptable, depending on the circumstances, and those above ± .21 are likely to be useful in most contexts. The validity coefficients for all expected scale relationships exceed ± .11, and in most cases, ± .21, in the following scale intercorrelation tables (see Tables 12–14).

As a rule, there is an inverse relationship between sample size and these critical values. For validity tests with test groups larger than 32, the critical value decreases for interpreting validity results as "acceptable" and "useful." For example, by doubling the size of a study from 32 to 64 participants, the value for acceptable validity decreases from ± .11 to ± .076. In fact, the DOL cautions employers in interpreting validity results involving studies with extremely large samples, in effect, lowering the suggested values for acceptable validity to minuscule values. Even so, with test groups around 2,000 participants, we adopt the guidelines provided for the much-smaller sample size of 32 in interpreting the validity coefficients for *PXT Select™.* This means *PXT Select* adheres to the most stringent, conservative interpretation of validity coefficients, no matter how large the validity test group is in the analyses reported below.

Thinking Style

A test group of 2,156 respondents completed the *PXT Select* Thinking Style section. Upon assessment completion, correlations among their results for the four cognitive scale measures were calculated. The scales on the *PXT Select* Thinking Style section demonstrate good convergent and divergent validity, as shown by the values listed in Table 12. These results demonstrate strong correlations that are likely to be useful in most contexts between the two verbal scales and between the two numerical scales. The correlations are significant at the .01 level, 2-tailed, indicating that the correlations provide strong support for our expectation that certain scales should be related.

Table 12. Thinking Style Scale Intercorrelations

	Composite Score	Verbal Skill	Verbal Reasoning	Numerical Ability	Numeric Reasoning
Verbal Skill	0.502				
Verbal Reasoning	0.743	0.216			
Numerical Ability	0.783	0.207	0.42		
Numeric Reasoning	0.832	0.188	0.537	0.673	
Correlations are significant at the 0.01 level (2-tailed).					

Behavioral Traits

A test group of 2,156 respondents completed the *PXT Select™* Behavioral Traits section. Upon assessment completion, correlations among their results for the nine behavioral scale measures were calculated. The scales on the *PXT Select* Behavioral Traits section demonstrate good convergent and divergent validity, as shown by the values listed in Table 13. These results demonstrate strong correlations that are likely to be useful in most contexts between different, but related, scales. The correlations are significant at the .01 level, 2-tailed, indicating that the correlations provide strong support for our expectation that certain scales should be related.

Table 13. Behavioral Traits Scale Intercorrelations

	PACE	ASSE	SOC	CONF	OUTL	ACCO	IND	JUDG	DEC
Assertiveness	0.422								
Sociability	0.336	0.325							
Conformity	-0.243	-0.338	0.151						
Outlook	-0.008**	-0.085	0.218	0.345					
Accommodation	-0.266	-0.552	0.05*	0.56	0.32				
Independence	0.187	0.271	-0.16	-0.618	-0.294	-0.394			
Judgment	-0.224	-0.107	-0.064	0.153	0.01**	0.105	-0.188		
Decisiveness	0.351	0.25	0.081	-0.137	0.063	-0.168	0.2	-0.242	
Correlation is significant at the 0.01 level (2-tailed). *Correlation is significant at the 0.05 level (2-tailed). ** Correlation is not significant.									

Interests

A test group of 2,156 respondents completed the *PXT Select™* Interests section. Upon assessment completion, correlations among their results for the six interest scale measures were calculated. The scales on the *PXT Select* Interests section demonstrate good convergent and divergent validity, as shown by the values listed in Table 14. These results demonstrate strong correlations that are likely to be useful in most contexts between different, but related, scales. The correlations are significant at the .01 level, 2-tailed, indicating that the correlations provide strong support for our expectation that certain scales should be related.

Table 14. Interests Scale Intercorrelations

	Enterprising	Financial/ Admin	People Service	Technical	Mechanical	Creative
Financial/Admin	-0.174					
People Service	0.485	0.046*				
Technical	-0.136	0.383	-0.202			
Mechanical	-0.139	-0.118	-0.176	0.156		
Creative	0.020**	-0.071	0.271	-0.286	-0.117	

Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).
**Correlation is not significant.

Correlations with Other Assessments

Thinking Style

An assessment measuring approximately the same aspects of verbal and numerical reasoning as *PXT Select* is the Adapt-*g* reasoning assessment. Adapt-*g* is an adaptive cognitive test applicable for all levels of participants' reasoning ability. It is used in employment settings, to determine employees' potential for promotion and whether they possess the right abilities for placement in a certain role. The assessment covers three areas of reasoning ability: Verbal Reasoning, Numerical Reasoning, and Abstract Reasoning. Results on the three measures are combined to form a General Mental Ability result.

A test group of 281 respondents completed both the *PXT Select* Thinking Style section and the Adapt-*g* assessment. Correlations among their results on the five *PXT Select* cognitive scales and the four Adapt-*g* scales were calculated. The scales on the *PXT Select* Thinking Style demonstrate good concurrent validity, as shown by the values listed in Table 15. All correlations are significant at the .01

level, 2-tailed, providing strong support for our expectation that certain scales should be related. In sum, the strongest scale relationship is between the Adapt-*g* overall General Mental Ability and the *PXT Select™* Composite Score. Other strong relationships are present between the numerical and verbal scales of both assessments.

Table 15. Correlations Between Thinking Style and Adapt-g Scales

	Adapt-g Scales			
PXT Select Scales	Verbal Reasoning	Numerical Reasoning	Abstract Reasoning	General Mental Ability
Composite Score	.68	.75	.66	.80
Verbal Skill	.40	.27	.31	.35
Verbal Reasoning	.68	.61	.58	.69
Numerical Ability	.44	.65	.50	.64
Numeric Reasoning	.59	.73	.64	.77

Behavioral Traits

An assessment measuring some of the same aspects of behavior as *PXT Select* is the *NEO™ Personality Inventory–3* (NEO™-PI-3). The NEO-PI-3 is a 240-item assessment measuring the Five-Factor Model of Personality: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience (McCrae & Costa, 2010). It is designed for a wide range of uses, including business and industrial settings and for psychological research. It also provides scores on six additional aspects of each of the five main scales, for a total of 30 additional measures. These areas include functional expressions of each of the five scales, such as Impulsiveness, Compliance, Deliberation, Assertiveness, and Trust. These areas, along with the five main factors, correspond well to the various behavioral scales of *PXT Select.* Our criterion of expected relationships among the nine *PXT Select* behavioral scales and the 35 NEO-PI-3 scales is supported by the correlations shown in the following tables (see Tables 16–24).

A test group of 300 respondents completed both the *PXT Select* Behavioral Traits section and the NEO-PI-3 assessment. Correlations among their results on the nine *PXT Select* behavioral scales and the 35 NEO-PI-3 scales were calculated. The scales on the *PXT Select* Behavioral Traits demonstrate good concurrent validity, as shown by the values listed in the following tables (see Tables 16–24). All correlations are significant at the .01 level, 2-tailed, except where noted, providing strong support for our expectation that certain scales should be related.

The five strongest correlations with the Pace scale are listed in Table 16.

Table 16. Strongest Correlations Between the Pace Scale and the NEO-PI-3

NEO-PI-3 Scale	Pace
Agreeableness	-.366
Compliance	-.277
Extraversion	.271
Assertiveness	.321
Activity	.525

The scales listed in Table 16 reflect the overall lively, active nature that is measured by the Pace scale. In addition, higher results on Pace were negatively correlated with Agreeableness and Compliance. This suggests that more urgent individuals tend to be less agreeable and compliant (as measured by the NEO™-PI-3) than those who are less urgent. This reflects the less patient, more aggressive aspects of the Pace scale.

The five strongest correlations with the Assertiveness scale are listed in Table 17.

Table 17. Strongest Correlations Between the Assertiveness Scale and the NEO-PI-3

NEO-PI-3 Scale	Assertiveness
Modesty	-.469
Agreeableness	-.464
Activity	.406
Extraversion	.446
Assertiveness	.750

Not surprisingly, the scales listed in Table 17 reflect the assertive, outgoing, and active nature that is measured by the Assertiveness scale. In addition, higher results on Assertiveness were negatively correlated with Agreeableness and Modesty. This suggests that more forceful individuals tend to be less agreeable and modest (as measured by the NEO-PI-3) than those who are more unassuming. This reflects the less humble, more dynamic, and outspoken aspects of the Assertiveness scale.

The five strongest correlations with the Sociability scale are listed in Table 18.

Table 18. Strongest Correlations Between the Sociability Scale and the NEO-PI-3

NEO-PI-3 Scale	Sociability
Positive Emotions	.326
Assertiveness	.406
Extraversion	.501
Gregariousness	.506
Warmth	.531

The scales listed in Table 18 reflect the overall warm, outgoing, and extraverted nature that is also measured by the Sociability scale. Those with a more outgoing nature tend to also be warm, open to conversation, and inclined to seek out opportunities for sharing and collaboration. These findings support the close relationship between the sociable measures of the NEO™-PI-3 and the Sociability scale.

The five strongest correlations with the Conformity scale are listed in Table 19.

Table 19. Strongest Correlations Between the Conformity Scale and the NEO-PI-3

NEO-PI-3 Scale	Conformity
Assertiveness	-.438
Angry Hostility	-.249
Modesty	.318
Agreeableness	.472
Compliance	.541

The scales listed in Table 19 reflect the overall agreeable, humble, and compliant nature that would be characteristic of someone with a high Conformity result. In addition, those who are more assertive and exhibit behaviors associated with anger or hostility (as measured by the NEO-PI-3) tend to be far more strong-willed than others who are more compliant.

The five strongest correlations with the Outlook scale are listed in Table 20.

Table 20. Strongest Correlations Between the Outlook Scale and the NEO-PI-3

NEO-PI-3 Scale	Outlook
Angry Hostility	-.324
Neuroticism	-.192
Compliance	.249
Agreeableness	.301
Trust	.525

The scales listed in Table 20 reflect the overall trusting, agreeable, and somewhat compliant temperament typical of someone with a more trusting Outlook result. In addition, those who exhibit more anger, hostility, and neuroticism (as indicated by the NEO™-PI-3) tend to be far more skeptical on the *PXT Select™* Outlook scale than those who are more trusting.

The five strongest correlations with the Decisiveness scale are listed in Table 21.

Table 21. Strongest Correlations Between the Decisiveness Scale and the NEO-PI-3

NEO-PI-3 Scale	Decisiveness
Deliberation	-.269
Vulnerability	-.166
Neuroticism	-.158
Anxiety	-.139
Assertiveness	.169

Although there is no direct conceptual correlate to the Decisiveness scale on the NEO-PI-3, the scales listed in Table 21 indicate the deliberating and vulnerable nature of those who are more deliberate on the Decisiveness scale. Those who are more bold on the Decisiveness scale, however, score higher on the NEO-PI-3 Assertiveness scale. These results characterize a bold decision-making style associated with a slightly more aggressive nature. Others more deliberate in making decisions may be more prone to worrisome thoughts, perhaps capturing the risk-aversion common for more deliberate individuals.

The five strongest correlations with the Accommodation scale are listed in Table 22.

Table 22. Strongest Correlations Between the Accommodation Scale and the NEO-PI-3

NEO-PI-3 Scale	Accommodation
Assertiveness	-.399
Angry Hostility	-.305
Modesty	.294
Agreeableness	.504
Compliance	.628

Table 22 shows a very strong tendency for those who are more agreeable on the Accommodation scale to be more compliant, agreeable, and modest as measured by the NEO™-PI-3. Taken together, these attributes reflect someone with a humble disposition who values harmony. In contrast, those with more steadfast perspectives on the Accommodation scale tend to also be more assertive and regard situations or others with more anger or hostility than those who are more agreeable.

The five strongest correlations with the Independence scale are listed in Table 23.

Table 23. Strongest Correlations Between the Independence Scale and the NEO-PI-3

NEO-PI-3 Scale	Independence
Agreeableness	-.250
Modesty	-.222
Compliance	-.205
Trust	-.153
Straightforwardness	-.148

Although there is no direct conceptual correlate to the Independence scale on the NEO-PI-3, the scales in Table 23 correlating with the Independence scale indicate that more reliant individuals also have a somewhat more agreeable, humble, compliant, and trusting nature than those who are more autonomous. This suggests that those who exhibit more autonomy are less inclined to go along with others for the sake of harmony, less restricted by rigid guidelines and instruction, and are more confident than those who prefer more guidance and instruction.

The five strongest correlations with the Judgment scale are listed in Table 24.

Table 24. Strongest Correlations Between the Judgment Scale and the NEO-PI-3

NEO-PI-3 Scale	Judgment
Impulsiveness	-.314
Feelings	-.281
Extraversion	-.248
Fantasy	-.223
Deliberation	.329

The scales shown in Table 24 related to the *PXT Select™* Judgment scale show that more factual individuals are also more deliberative. However, those who are more intuitive on the Judgment scale can be more impulsive and outgoing, place higher emphasis on feelings, and tend to focus more on possibilities than those who are more factual. This suggests a more deliberative, practical nature of those who are more factual, and a more open-minded perspective of those who are more intuitive.

Interests

PXT Select uses Holland's six constructs—Realistic, Investigative, Artistic, Social, Enterprising, and Conventional—to measure occupational interest. Holland's *Vocational Preference Inventory* (1985) and the body of supporting research have emerged as the standard for measuring one's work interest. The *PXT Select* Interests section is influenced by this classification for work-related interests, measuring an individual's level of interest in six areas:

PXT Select Interests	Holland's Constructs
1. Enterprising	Enterprising
2. Financial/Administrative	Conventional
3. People Service	Social
4. Technical	Investigative
5. Mechanical	Realistic
6. Creative	Artistic

These areas parallel those found in Holland's typology, which are widely accepted as important factors in job success. The names of the interest areas in *PXT Select* have been updated to make them more

relevant to current business and industry. The assessment identifies the respondent's top three occupational interests and describes their role in an individual's approach in the workplace.

To test the validity of the *PXT Select™* interest scales in measuring Holland's six constructs, a test group of 190 respondents completed the *PXT Select* Interests section and Holland's *Self-Directed Search* (Holland, 1994). Correlations among their results on the six *PXT Select* interest scales and the six *Self-Directed Search* scales were calculated. The scales on the *PXT Select* Interests section demonstrate good concurrent validity, as shown by the values listed in Table 25. All correlations are significant at the .01 level, 2-tailed, providing strong support for our expectation that the six interest scales do, in fact, measure the six areas of Holland's occupational interests typology. The results demonstrate that the *PXT Select* Interests section is closely related to the constructs of the *Self-Directed Search*.

Table 25. Correlations Between Interests and Self-Directed Search Scales

PXT Select	SDS Scales	Correlation Coefficient
Enterprising	Enterprising	.43
Financial/Administrative	Conventional	.50
People Service	Social	.44
Technical	Investigative	.51
Mechanical	Realistic	.65
Creative	Artistic	.61
All correlations significant at the .01 level.		

Summary of the Validation Results

Evaluation of the *PXT Select* assessment confirms adherence to the U.S. Department of Labor's three Principles of Assessment relating to reliability and validity. Reliability results upheld both scoring and assessment item measurement consistency. The reliability results meet, and in almost all cases exceed, DOL guidelines for interpreting reliability coefficients as acceptable rates of measurement consistency, with low rates of measurement error.

Convergent and discriminant validity was supported by favorable scale intercorrelation confirming expected *PXT Select* scale relationships. Further, evidence for validity was also provided by strong

correlations between *PXT Select™* and known standards of cognitive and behavioral measures, Adapt-*g* and NEO™-PI-3. The results presented in the Reliability and Validity sections meet and exceed the DOL guidelines for interpreting validity coefficients, confirming the usefulness of the assessment in the context and for the audience for which it was designed.

Performance Model Library

The *PXT Select* library of Performance Models was created to give human resource professionals a greater chance of success in their hiring efforts. Work-related performance ratings, encompassing actual performance metrics of various organizations, were used to create a number of library models. Other library models were created by leveraging U.S. Department of Labor data available through the government's data collection program at https://onet.rti.org. Additional information about job descriptions, titles, and job families may be found at onetonline.org.

Performance-based Models

To build each performance-based library model, incumbent employees' job titles, assessment results, and job performance ratings were analyzed. This analysis produced score ranges and interests reflecting those of effective performers in each type of job. To validate each model, the criterion-related concurrent validity coefficient (correlation) was calculated between job performance and the Overall Fit score produced by matching the original dataset of employees' assessment results to the library model. Most of the performance-based library models exhibited "very beneficial" validity according to the general guidelines for interpreting validity coefficients (U.S. Department of Labor, 2000, p. 3-10). Other models exhibited "likely to be useful" validity, demonstrating good utility and concurrent, criterion-related validity of the models for indicating potential job success.

O*Net-based Models

The National Center for O*Net Development, a U.S. Department of Labor initiative, describes worker characteristics as "enduring characteristics that may influence both work performance and the capacity to acquire knowledge and skills required for effective work performance" (National Center for O*Net Development, n.d., p. 1, as cited in Burrus, Jackson, Xi, & Steinberg, 2013, p. 7). *PXT Select* Thinking Style, Behavioral Traits, and Interests scales were mapped to the most recent O*Net dataset on required Abilities, Work Styles, and Interests for effective work performance. Further information on the development of the O*Net abilities, work styles, and interest constructs used in creating the O*Net Performance Models, as well as the sample group characteristics and survey and sampling

methodology, can be found in the research report, *Identifying the Most Important 21st Century Workforce Competencies: An Analysis of the Occupational Information Network (O*Net)* (Burrus, et al., 2013).

The DOL job requirement ratings for each position contain quantitative information about the level and the importance of each requirement that is typical for effective performance in the position. O*Net data level and importance ratings were standardized according to a normal distribution of scores that would represent those applicable for the U.S. population, and converted into values corresponding to *PXT Select™* STEN scores. Aspects of abilities, work styles, and interests found by the DOL that correspond to effective job performance are highly similar to the scales measured by *PXT Select.* So similar, in fact, as to be appropriate for the purposes of building Performance Models for use in determining job fit for respondents completing *PXT Select.* The mapping between O*Net constructs and *PXT Select* scales is shown in Tables 26–28.

Table 26. Correspondence Between Thinking Style and O*Net Abilities

PXT Select Thinking Style Scales	O*Net Abilities
1. Composite Score	Combination of Abilities
2. Verbal Skill	Written Comprehension
3. Verbal Reasoning	Deductive & Inductive Reasoning
4. Numerical Ability	Number Facility
5. Numeric Reasoning	Mathematical Reasoning

Table 27. Correspondence Between Behavioral Traits and O*Net Work Styles

PXT Select Behavioral Traits Scales	O*Net Work Styles
1. Pace	Achievement/Effort
2. Assertiveness	Leadership
3. Sociability	Social Orientation
4. Conformity	Dependability
5. Outlook	Concern for Others
6. Decisiveness	Initiative
7. Accommodation	Cooperation
8. Independence	Independence
9. Judgment	Analytical Thinking

Table 28. Correspondence Between Interests and O*Net Interests (Holland's Constructs)

PXT Select Interests Scales	O*Net Interests
1. Enterprising	Enterprising
2. Financial/Administrative	Conventional
3. People Service	Social
4. Technical	Investigative
5. Mechanical	Realistic
6. Creative	Artistic

To validate each O*Net-based Performance Model for indicating potential job success, *PXT Select™* results for respondents in various jobs and industries were compared to the O*Net-based Performance Model reflecting their position. The Overall, Thinking Style, Behavioral Traits, and Interests fit percentages were calculated for *PXT Select* respondents in the same position represented by each type of O*Net Performance Model. In general, there was good variability in Overall Fit results approaching a normal distribution for each model when compared to the entire test sample of *PXT Select* respondents. Further analysis showed higher values of fit percentages when results were compared to jobs corresponding to respondents' reported job title and industry. This provides evidence for the utility of the models in differentiating between attributes of individuals holding various jobs represented by various O*Net-based Performance Models.

Legal Issues: Fairness and Adverse Impact

The U.S. Department of Labor's Principle of Assessment about assessment fairness recommends that employers "use only assessment instruments that are *unbiased* and *fair* to all groups" (U.S. Department of Labor, 2000, p. 2-1). This is to prevent any chance of the assessment introducing discrimination to the selection process against any group of individuals who identify with a certain demographic group. Assessments that are fair and unbiased give everyone an equal opportunity to perform on the assessment, and, by extension, an equal opportunity for employment selection based in part on assessment results. We examined the two aspects of this Principle of Assessment—"fair" and "unbiased"—in different ways. For fairness, we ensured the equality of assessment results among various demographic groups who have completed *PXT Select* as applicants for a job. To test that the assessment is unbiased, we examined the selection rate of applicants across these same demographic groups. This is to ensure the assessment does not introduce adverse impact on the selection rate for any demographic group.

To evaluate adverse impact, an Overall Fit percentage of 70% was used as a cutoff. That is, in this hypothetical scenario, it can be assumed that candidates with percentages at or above 70% would be hired by a given company and those below this threshold would be rejected. This provides an objective method to evaluate how fit scores might impact selection ratios across different demographic groups. Please note that *we do not recommend using a rigidly defined cut-score when making actual hiring decisions.* It is critical to note that the fairness and unbiased nature of the assessment itself can be verified through studies like the ones reported here. However, the fairness of any part of an organization's selection process, or the process as a whole, cannot be verified by the publisher whose assessment is used as one part of the selection process. It is the organization's responsibility to maintain fairness throughout the entire selection process, and for every hiring evaluation, including the interview, qualification tests, screening procedures, and so forth. For this reason, we recommend that the assessment serve as no more than one-third of the ultimate hiring decision. This recommendation also extends to other types of personnel decisions, but is most critical in the hiring context. There are various forms of legislation, governing bodies, and governmental agencies involved in establishing and enforcing the rules around ensuring fairness in employment opportunity. For more specifics on these types of regulations and enforcing bodies, please see Chapter 2 of the U.S. Department of Labor's *Testing and Assessment: An Employer's Guide to Good Practices.*

Characteristics of the Sample Group

A sample group of 260 applicants completed the assessment as candidates for employment in various organizations. The applicants' age, gender, and ethnicity were represented in the approximate proportions of each group in the overall U.S. population, reflected in 2010 and 2015 U.S. Census Bureau statistics. Participants completed the assessments online, in an unproctored setting, identical to the conditions under which the *PXT Select™* assessment is administered to all respondents. Their assessment results were compared to a Performance Model reflecting the position to which they were applying, resulting in an Overall Fit percentage, as well as fit percentages for the Thinking Style, Behavioral Traits, and Interests sections. These overall and sectional fit percentages are used as the basis of comparison of assessment results. This is because fit percentages are the metrics indicating job fit, and as such, are used for determining fairness and any potential adverse impact across demographic groups.

Fairness and Adverse Impact

The DOL guidelines describe the most important aspects of the various types of legislation concerning equal employment opportunity and the use of assessments in hiring. One of the most important aspects concerns the fairness of assessments for different groups of individuals whose access to employment is protected for every part of the hiring process. The DOL recommends that employers should use professionally developed assessments that do not discriminate against any protected group. Applicable for employers with 15 or more employees, the protected groups described in Title VII of the Civil Rights Act of 1964 for the United States include both male and female gender groups, those who are age 40 and older, and people who identify with racial, religious, or ethnic groups.

According to various governmental regulations, if an assessment is to be used as part of a workforce decision process, including, most consequentially, a selection decision for employment, the assessment may not adversely impact any demographic group with which a respondent identifies. However, the final personnel decision, and in particular, a candidate selection decision, involves many different aspects of the entire hiring process that are outside the assessment. Therefore, we need some way to determine the rate of impact the assessment has on the hiring process based on demographic group.

For *PXT Select™,* the critical value in the workforce decision-making procedure is the Overall Fit. This is the key indicator of the degree to which a respondent will likely fit the position represented by the Performance Model. Decision makers may reference this Overall Fit value as an estimation of the respondent's potential in the job. In referencing the Overall Fit result to make decisions, and knowing that higher degrees of potential job fit are indicated by higher values of the Overall Fit result, we want to ensure that similar degrees of Overall Fit results are found for all demographic groups. Assuming that higher values of Overall Fit reflect positively on the respondent, perhaps contributing to a favorable workforce decision, relatively high values of Overall Fit are used to define "selection" for the purposes of these analyses. In particular, we identify Overall Fit values exceeding 70% as sufficiently high degrees of job fit that could reflect favorably on the candidate.

Government guidelines, and the DOL Principle of Assessment related to adverse impact, mandate that assessment results such as this metric should not adversely impact the placement decision for individuals of a certain demographic group. In order to determine that *PXT Select* does not adversely impact any workforce decision for any demographic groups, we analyze the selection rates for each demographic group to ensure that they are sufficiently equivalent.

The degree of selection rate equivalence, and the value at which sufficient equivalence is reached, is established by the U.S. government's Uniform Guidelines on Employee Selection Procedures (1978) "Four-Fifths Rule." It states, "A selection rate for any race, sex, or ethnic group which is *less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate* will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by the Federal enforcement agencies as adverse impact" (Equal Employment Opportunity Commission, et al.,1978, Section 4, Article D, emphasis added).

When establishing the fairness of the assessment with respect to selection rates, and its compliance with the Four-Fifths Rule, we must define "selection rate" so as to isolate the influence of the assessment itself in determining the final selection rate of candidates among demographic groups. Because high degrees of candidate fit exceeding 70% could conceivably be considered favorable for any respondent, we define "selection rate" as an Overall Fit result that is 70% or above. If a respondent's scores are matched to the Performance Model, resulting in an Overall Fit of 70% or above, we say they are "selected" for the position, as identified by the assessment. This does not necessarily mean they were ultimately selected for the job as an employee of the client organization using the assessment—there are many other factors influencing the final decision outside the assessment. It simply means the assessment has indicated a favorable degree of job fit for the candidate, predicting a good chance that the candidate, if selected, will be successful in the role.

To establish whether adverse impact would be introduced to the workforce decision process by *PXT Select™,* we compare the impact ratios for each group within the demographics of gender, age, and ethnicity. To calculate the impact ratio, we first calculate the percentage of respondents in each demographic group whose Overall Fit result is 70% or above. This gives the selection rate for each demographic group. Then, we identify which group has the highest selection rate. According to the Four-Fifths Rule, all other demographic groups must have a selection rate that is at least four-fifths, or 80%, of the group with the highest selection rate. This ratio is the impact ratio—the ratio of the selection rate of each demographic group to the selection rate of the highest-selected group. The impact ratio must exceed .80 to satisfy the Four-Fifths Rule.

For gender, t-tests revealed statistically significant differences ($p < .05$) in three of the four analyses: Overall Fit score, Thinking Style Fit score, and Behavioral Traits Fit score.

The gender group with the highest selection rate is the Female group. The impact ratio for gender, then, is calculated by dividing the Male group's selection rate by the Female group's selection rate. If the impact ratio exceeds .80, then the Four-Fifths Rule is satisfied, and the assessment is considered to introduce no adverse impact to the workforce decision process with respect to gender.

Analyses revealed that the gender impact ratio is .84, exceeding .80, as required by the Four-Fifths Rule in establishing the impact of selection rates of the assessment between gender groups. As far as the assessment influences the overall workforce decision process, according to the Four-Fifths Rule, it introduces no adverse impact to the process on the basis of gender. Taken together, although some variation in fit scores was found between the gender groups, it is expected to have no practical impact on the hiring process.

For age, t-tests first compared the fit scores of those aged 40 and below with those aged 41 and above. No statistically significant differences were found at $p < .05$.

The age group with the highest selection rate is the 40-and-younger age group. The impact ratio for age, then, is calculated by dividing the 41-and-older age group's selection rate by the 40-and-younger group's selection rate. If the impact ratio exceeds .80, then the Four-Fifths Rule is satisfied, and the assessment is considered to introduce no adverse impact to the workforce decision process with respect to age.

Analyses indicate that the age impact ratio is .95, well exceeding .80, as required by the Four-Fifths Rule in establishing the impact of selection rates of the assessment between age groups. As far as the assessment influences the overall workforce decision process, according to the Four-Fifths Rule, it introduces no adverse impact to the process on the basis of age.

For ethnicity, ANOVAs (analysis of variance) were first performed comparing the mean fit scores of four ethnic groups: Asian or Pacific Islander, African-American, Caucasian, and Hispanic. The highest F values were for the Interests section ($F(4, 246) = 2.950$; $p < .05$) and the Thinking Style section ($F(4, 246) = 2.40$; $p < .05$). F values for the Overall Fit and Behavioral Traits section were not statistically significant at $p < .05$.

The ethnic group with the highest selection rate is the Hispanic group. The impact ratios of the other ethnic groups, then, are calculated by dividing each of their selection rates by the Hispanic group's

selection rate. If each group's impact ratio exceeds .80, then the Four-Fifths Rule is satisfied, and the assessment is considered to introduce no adverse impact to the workforce decision process with respect to ethnicity.

Three impact ratios were calculated, for each demographic group that was selected at a lower rate than the highest-selected ethnicity, the Hispanic group. The lowest of these impact ratios was .90, with each far exceeding the .80 threshold, as required by the Four-Fifths Rule in establishing the impact of selection rates of the assessment between ethnic groups. As far as the assessment influences the overall workforce decision process, according to the Four-Fifths Rule, it introduces no adverse impact to the process on the basis of ethnicity. Taken together, although some variation in fit scores was found among the ethnic groups, it is expected to have no practical impact on the hiring process.

Summary of Fairness and Adverse Impact Findings

The impact ratio analysis results demonstrate a lack of adverse impact for any one demographic group that completes the *PXT Select™* assessment. This process confirms that *PXT Select* provides all respondents an equal opportunity for employment on the basis of the assessment results, at least for the part of the employment opportunity that is informed by *PXT Select* results.

Even though the assessment results used for any personnel decision should account for no more than one-third of the final decision, it is incumbent on the assessment publisher to ensure that the assessment portion of the decision should not introduce adverse impact to the overall decision process for any demographic group. To this end, *PXT Select* does not result in selection rates, as defined by relatively high degrees of fit percentages, that vary meaningfully as a result of respondents' gender, age, or ethnic group. *PXT Select* is in full compliance with the DOL Principle of Assessment suggesting employers only use assessments that are unbiased and fair to all groups, as well as all legislation pertaining to equality, fairness, and adverse impact of employment assessments.

References

Ackerman, P.L., & Heggestad, E.D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*, 219–245.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Burrus, J., Jackson, T., Xi, N., & Steinberg, J. (2013). *Identifying the most important 21st century workforce competencies: An analysis of the Occupational Information Network (O*Net)* (ETS Research Report). Princeton, NJ: ETS.

Cronbach, L.J. (1984). *Essentials of psychological testing* (4th ed.). San Francisco: Harper & Row.

Equal Employment Opportunity Commission, 29 Fed. Reg. pt. 1,607, Department of Labor Office of Federal Contract Compliance Programs, 41 Fed. Reg. pt. 60-3, Department of Justice, 28 Fed. Reg. 50.14, Civil Service Commission, 5 Fed. Reg. 300.103(c) (1978, Aug. 25).

Holland, J.L. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6,* 35–45.

Holland, J.L. (1985). *Manual for the vocational preference inventory.* Odessa, FL: Psychological Assessment Resources.

Holland, J.L. (1994). *The self-directed search*. Odessa, FL: Psychological Assessment Resources.

McCrae, R.R., & Costa, P.T., Jr. (2010). *NEO inventories: Professional manual*. Odessa, FL: Psychological Assessment Resources.

National Center for O*Net Development. (n.d.). *The O*Net content model: Summary outline.* Retrieved from http://www.onetcenter.org/dl_files/ContentModel_Summary.pdf

Parsons, F. (1909). *Choosing a vocation*. Boston, MA: Houghton.

Rounds, J.B., & Tracey, T.J. (1990). From trait-and-factor to person-environment fit counseling: Theory and process. In W.B. Walsh & S.H. Osipow (Eds.), *Career counseling: Contemporary topics in vocational psychology* (pp. 1–44). Hillsdale, NJ: Erlbaum.

U.S. Department of Labor, Employment and Training Administration. (2000). *Testing and assessment: An employer's guide to good practices*. Washington, DC: U.S. Printing Press.

Uniform guidelines on employee selection procedures. (1978). 43 F.R. 38295.

Williamson, E.G. (1939). *How to counsel students: A manual of techniques for clinical counselors*. New York, NY: McGraw-Hill.