Technical Summary & Test Review

Reviewer/Psychometrician	Gordon Goodwin
Date of Current Review	2/15/2021
Date of Previous Review	N/A
Instrument Name	[redacted] in Python Programming
Current Version	[redacted]
Retired Versions	[redacted]
Test Publisher	[redacted]
Subject Matter Experts	[redacted]
Current Version Date of Publication	

Assessment Summary

General Overview	
The [redacted] ([redacted]) measures associate-level (intermediate) proficiency in the Python programming language (Van Rossum & Drake Jr., 1995). The [redacted] certification exam can be taken as a stand-alone certification exam or following the completion of a series of open-access courses. Examinees who receive a passing score are awarded an Associate-level certification. The exam consists of 5 topical sections pertaining to intermediate-level fundamentals of Python programming. [redacted] Examinees may take the assessment multiple times. Examinees receive a random sampling of one of three different versions for each of the 40 topical items, and the items they receive are presented in random order within each section.	
Content Domain(s)	Ability – Python language programming
Intended/Main Area of Use	Educational
Intended Population	International/Open/No prerequisites
Scales/Topical Sections	Single Scale with 5 Topical Sections 1) Modules & Packages 2) Exceptions 3) Strings 4) Object-Oriented Programming (OOP) 5) Miscellaneous
Delivery Channel	Computerized: • [redacted] Testing Centers • [redacted] proctored
Administration/Oversight	Timed & Controlled (proctored)
Test Duration	75 minutes total • Tutorial/NDA: 10 minutes • Exam: 65 minutes

Item Format	Multiple Choice (A/B/C/D) variations:
	• 18 single response items
	  o Dichotomous: no credit/full credit
	• 22 double response items (2 answers)
	  o Partial credit: 0 credit/half credit/full

Assessment Length/Structure	40 items total across 5 topical sections 1) Modules & Packages: 6 items 2) Exceptions: 5 items 3) Strings: 8 items 4) Object-Oriented Programming (OOP): 12 items 5) Miscellaneous: 9 items
Test Bank Size	120 items total • Each of the 40 topical items have 3 versions ○ Ex: Q1_V1, Q1_V2, Q1_V3
Item Distribution/Sampling Method	Random-random • For each of the 40 items the examinee receives, one of the three possible versions is randomly sampled • Within each of the 5 sections, the order of the items is randomly generated
Cut Score	70% Pass/Fail

Scoring Method	• Total of 100 points possible
	• Differential scoring/weighting based on item difficulty
	  ○ Items worth either 2 or 4 points total
	  ○ Weights based on SME guidance
	• Computerized scoring
	  ○ Examinee enters responses, scores calculated by computer

Feedback	Score report post-administration
Navigation Format	Linear, with ability to return to items

Demands on Examinee & Accommodations	• Vision
	  ○ Zoom & Color accommodations available
	• Speed Reading
	  ○ Time accommodation available
	• English Proficiency
	  ○ English only

Costs/Fees	• [redacted] • Free practice exam via open-access

Topical Section/Subscale Content Domains	
Modules & Packages	• import variants; advanced qualifiying for nested modules • dir(); sys.path variable • math: ceil(), floor(), trunc(), factorial(), hypot(), sqrt(); random: random(), seed(), choice(), sample() • platform: platform(), machine(), processor(), system(), version(), python_implementation(), python_version_tuple() • idea, __pycache__, __name__, public variables, __init__.py

	• searching for modules/packages; nested packages vs directory tree
Exceptions	• except, except:-except; except:-else:, except (e1,e2) • the hierarchy of exceptions • raise, raise ex, assert • event classes, except E as e, arg property • self-defined exceptions, defining and using
Strings	• ASCII, UNICODE, UTF-8, codepoints, escape sequences • ord(), chr(), literals • indexing, slicing, immutability • iterating through, • concatenating, multiplying, comparing (against strings and numbers) • in, not in • .isxxx(), .join(), .split() • .sort(), sorted(), .index(), .find(), .rfind()
Object-Oriented-Programming	• ideas: class, object, property, method, encapsulation, inheritance, grammar vs class, superclass, subclass • instance vs class variables: declaring, initializing • __dict__ property (objects vs classes) • private components (instance vs classes), name mangling • methods: declaring, using, self parameter • instrospection: hasattr() (objects vs classes), __name__, __module__, __bases__ properties • inheritance: single, multiple, isinstance(), overriding, not is and is operators • inheritance: single, multiple, isinstance(), overriding, not is and is operators • constructors: declaring and invoking • polymorphism • __name__, __module__, __bases__ properties, __str__() method • multiple inheritance, diamonds
Miscellaneous	• list comprehension: if operator, using list comprehensions • lambdas: defining and using lambdas, self-defined functions taking lambda as as arguments; map(), filter(); • closures: meaning, defining, and using closures

	• I/O Operations: I/O modes, predefined streams, handles; text/binary modes open(), errno and its values; close() .read(), .write(), .readline(); readlines() (along with bytearray())

Evaluation Process & Findings

Evaluation Process General Overview
Validation of the [redacted] was conducted in alignment with the prescriptive guidance regarding educational and psychological assessment practices put forth in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME), *European Test User Standards* (EFPA, EAWOP) and the *European Test Review Model* (EFPA, EAWOP). The evaluation process consisted of a thorough review of all available evidence gathered during the design, development, and implementation of the [redacted]esting program, and involved an iterative collaboration between the test publisher, subject matter experts, and psychometrician(s). Further, data gathered from a field test of [redacted]espondents was also comprehensively reviewed. Consequently, validation of the [redacted] assessment and testing practices was carried out in a manner consistent with applicable industry best practices, ethical standards, and prominent research literature. This iterative process of collaboration between publisher, subject matter experts, and psychometrician and comparison of the evidence to applicable standards provided the basis for the findings and recommendations.

Item Response Theory (IRT): Verifying Unidimensionality
In comparison to traditional fixed-form exams, the [redacted] utilizes a "random-random" sampling procedure to randomly sample one of three versions of each of 40 items from the 120-item test bank. As such, no fixed-form versions of the [redacted] exist, which assists in preventing cheating and piracy. Consequently, item-level analyses were conducted under the guiding framework of Item Response Theory (IRT). In comparison to classical test theory (CTT), IRT is considered as the standard, if not preferred, method for conducting psychometric evaluations of new and established measures (Embretson & Reise, 2000; Fries et al., 2005; Lord, 1980; Osteen, 2010; Ware et al., 2000). At a high level, IRT is based on the premise that only two elements are responsible for a person's response on any given item: the person's ability (or abilities), and the characteristics of the item (Bond & Fox, 2001; Osteen, 2010). Development and validation of the [redacted] entailed the use of a unidimensional IRT model based on the premise that correlations among responses to test questions can be explained by a single underlying trait (i.e. Python proficiency/ability). While traits/abilities like Python proficiency are complex and represent many different constituent skills and facts that are combined in specific ways, the claim of unidimensionality is that these components work together to manifest a coherent whole. Although the test is structured around five topical sections, this was done to provide adequate domain sampling rather than to measure different traits. While individuals may have strengths and weaknesses with respect to the topical sections on a unidimensional test, any *systematic* relationship *among* those topical sections should be explained by the effect of the *single* latent trait or ability (Python proficiency) upon the examinees' item responses. In alignment with the literature standard, unidimensionality was evaluated (and confirmed) through use of a confirmatory factor analysis (CFA) model and review of goodness of fit statistics (RMSEA, CFI, TLI).

Item Response Theory (IRT): Model Overview

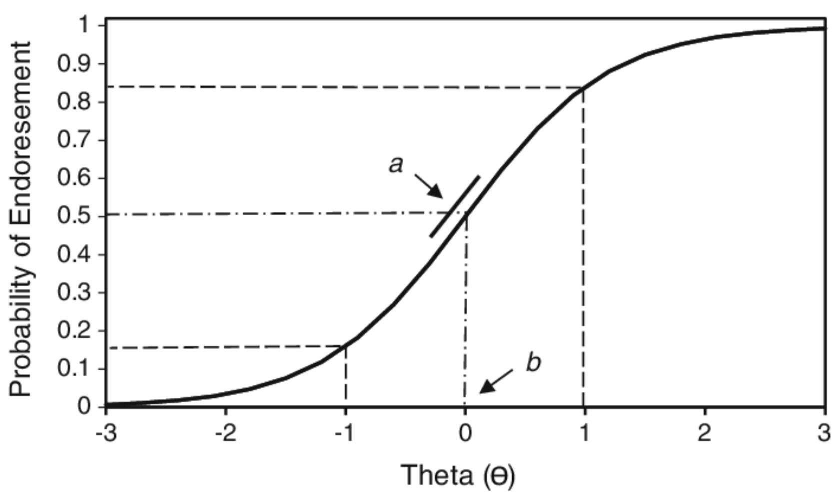
At a basic level, IRT models estimate mathematical equations in order to model the relationship between an examinee's probability of correctly responding to an item and their ability level. The basic unit of an IRT model is the Item Characteristic Curve (ICC), shown below, which estimates the probability of a given response based on a person's level of latent ability, wherein the shape and location of the curve is determined by the item characteristics estimated by the model parameters. While there are a variety of different forms an IRT model can take, IRT models of the form utilized for this evaluation assume the probability of a given response is a function of the person's *ability* (theta θ), the *difficulty of the item* (*b*), and the *discrimination of the item (a)*.

Specifically, the person-level ability level (θ) is calculated for each respondent on the basis of their overall test performance, with an ability value of ($\theta = 0$) representing an individual of *average* ability. Using this ability scale, the difficulty parameter (*b*) for each item then states the ability level required in order for a respondent to have a 50% probability of endorsing that item correctly. Consequently, for items with higher difficulty parameters (any positive value of $b > 0$) , only the examinees with above-average abilities ($\theta > 0$) will have a 50% probability of getting the item correct. For lower-difficulty parameters ($b < 0$), examinees with below average ability levels ($\theta < 0$) still have a 50% or greater chance of answering the item correctly.

The discrimination parameter (*a*) measures the differential capability of an item, such that a high discrimination parameter value (*a*) suggests the item differentiates well amongst subjects. Put simply, a high discrimination parameter value (*a*) means that the probability of a correct response increases rapidly as the underlying ability level increases, and a low discrimination parameter value means that the probability of getting a correct response on the item does not increase rapidly as the ability level increases. Items with high discrimination parameters (steep curves) are desirable in that a given examinee's response will be more *informative* about their underlying ability value. In contrast, for items with low discrimination parameters (shallow curves), subjects' responses aren't as informative about their underlying ability level because the probability of getting a correct response is relatively constant across ability levels.

The ICC plot provides a visual representation of the item characteristics or parameters estimated by the model. As seen in the exemplar ICC plot below, the difficulty parameter (*b*) governs the side-to-side location of the curve along the ability (θ) scale, with this particular plot representing an item of average difficulty ($b = 0$). While no specific estimate for the discrimination parameter (*a*) is provided for the ICC plot below, this item appears to differentiate reasonably well given that the curve retains its sigmoid shape and is not shallow.

Figure 1. Exemplar Item Characteristic Curve (ICC) for Dichotomous Items



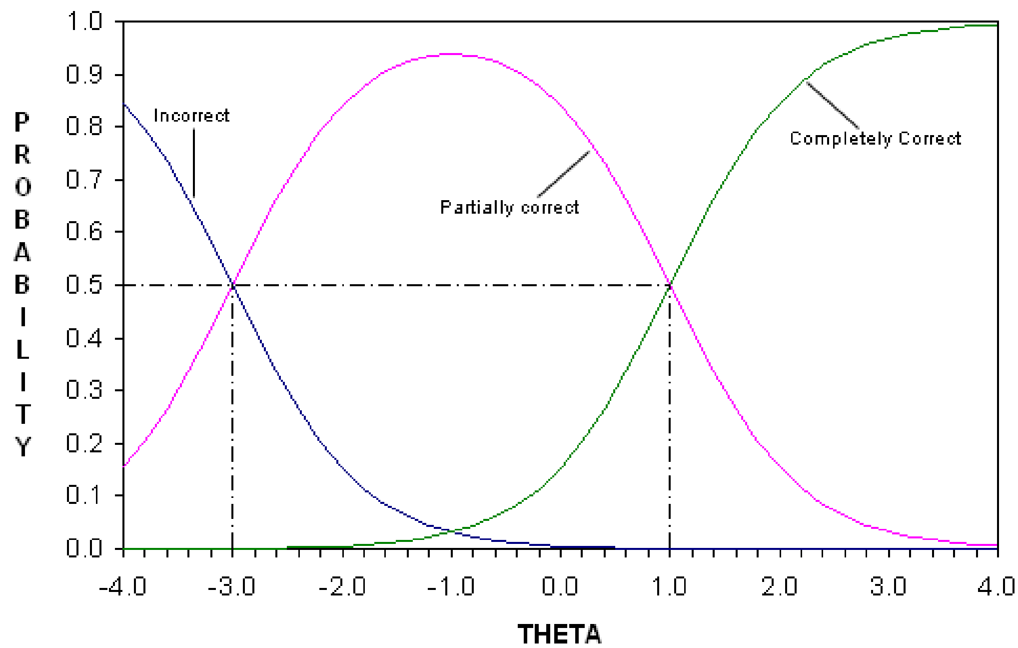
Generalized Partial Credit Model

Evaluation of the [redacted] was conducted utilizing a specific form of IRT model referred to as the *generalized partial credit model* ([GPCM]; Muraki, 1992), which allows for a mixture of dichotomous items (where a response is either completely right or wrong) and polytomous items (where examinees can receive partial credit for a partially-correct response). As seen below, the ICC plot takes a slightly more complex form with three separate curves for items that allow for partial credit, wherein the first curve represents the probability of getting zero credit, the second displays the probability of getting partial credit, and the third represents the probability of receiving full credit.

As such, under the GPCM model, there are two difficulty parameters estimated (b_1 and b_2) for items where partial credit is possible, one each for the partial credit and full credit curves respectively. This allows one difficulty parameter (b_1) to estimate the ability level required to have a 50% chance of crossing the threshold from receiving zero credit to half-credit, and another difficulty parameter (b_2) to estimate the ability level required to have a 50% chance of crossing the threshold from receiving half-credit to full-credit.

As before, the GPCM includes a discrimination parameter (a) that measures the differential capability of the item. Visually, the discrimination slope parameter (a) again manifests as the steepness or shallowness of the ICC plot, the first difficulty parameter (b_1) the side-to-side location of the partial-credit probability curve along the ability (θ) scale, and the second difficulty parameter (b_2) governs the side-to-side location of the full credit probability curve along the ability (θ) scale. A review of the GPCM parameters is provided below.

Figure 2. Exemplar Item Characteristic Curve (ICC) for Polytomous Items

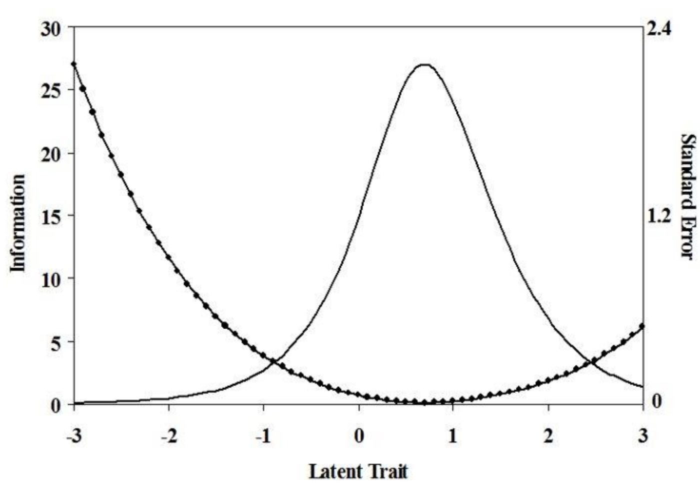


*In the above plot, b_1 = -3.0 marks the ability level at which examinees cross the threshold from zero credit to having a 50% probability of getting partial credit, and b_2 = +1.0 marks the ability level at which examinees cross the threshold from partial to having a 50% probability of receiving full credit.

Generalized Partial Credit Model Parameters		
b_1	Difficulty parameter for partial credit threshold	Ability (θ) level where the probability of moving from getting zero to partial credit is 50%
b_2	Difficulty parameter for full credit threshold	Ability (θ) level where the probability of moving from getting partial to full credit is 50%
a	Discrimination parameter	The slope of the curve at the difficulty location *b*, describes how well the item differentiates ability
θ	Person-level ability parameter	Standardized measure of examinee ability level, where 0 = average ability, based on subject's performance on the overall assessment

Item Information Function (IIF)
The information provided by an item and a test can be evaluated in an IRT model by using the *Item Information Function*, denoted as the IIF or as I(θ). The information for an item is essentially an index of how *precise or accurate* the item is over the range of ability levels (θ). If an item is very precise and accurate for individuals of a given ability level, then the item is very *informative* regarding that ability level. The Item Information Function plot basically provides a visual representation of this, such that the highest point on the IIF curve corresponds to the ability level for which the item is *most informative*. In addition, the *peakedness* of the IIF plots is also useful in that items with steep, narrow, peaked IIF curves denote that the item is highly informative over a specific range of ability. In contrast, shallow, less-peaked IIF curves denote items where a lesser amount of information is spread out over a wider range of ability levels. While the *Item Information Function* (IIF) represents the range of ability levels that each individual item is most informative over, the *Test Information Function* (TIF) represents the range of ability levels that the *test as a whole* is most informative over and functions most effectively. Just as the Item Information Function is related to how precise a given individual item is at different ability levels, the Test Information Function is related to how precise the *test* is across different ability levels. This overall accuracy and precision is indexed through the inverse of the *Standard Error of θ*, which simply quantifies the expected error for any estimate along the range of ability (θ) levels. In practical terms, when the TIF curve is concentrated over a below-average ability level ($\theta < 0$), as is the case with [redacted], the test is most effective and provides estimates with lowest standard error for individuals with lower ability levels. When the TIF is concentrated (peaked) over higher ability levels ($\theta > 0$), as is the case in Figure 2 below, this indicates the test as a whole is most effective at evaluating above-average ability levels.

Figure 2. Exemplar Test Information Function (TIF)



* In the above plot, the TIF is plotted against the Standard Error, which visually represents the inverse representation between information and error of measurement. The test is most effective over the range of ability levels where the standard error is lowest (in this case above average ability levels)

Applying IRT to [redacted]
After verifying the appropriateness of assuming a unidimensional underlying ability/trait, each of the 120 items in the [redacted] test bank were analyzed using a GPCM model. Specifically, difficulty and discrimination parameters, as well as ICC and IIF plots, were estimated and reviewed for all 120 items. This entailed the following process: 1) Difficulty (*b*) and discrimination (*a*) parameters were reviewed first at the item-version level, such that for each of the 40 topical items, parameter estimates were compared for the three available versions in order to ensure equivalency and fairness across the versions of each item. 2) After establishing consistency and general fairness across versions, differential item functioning (DIF) and measurement invariance were evaluated with respect to the available demographic variables (gender proxy and nationality). Also of note, a sensitivity review was conducted prior to the psychometric evaluation process, during which items deemed culturally insensitive or inappropriate to minority groups were removed from the test bank. 3) Having confirmed consistency and fairness in the context of both the item-version distribution and the across-groups measurement structure, parameter estimates for all 120 items were reviewed in order to identify items with extreme difficulty values ($b < -3$, $b > +3$) and/or low discrimination values for prospective removal. 4) ICC and IIF plots were reviewed for the remaining items to ensure that they individually and collectively represented a reasonable coverage across a diverse range of ability levels, and were particularly informative across the ability levels of interest ($b < 0$). 5) The TIF plot was reviewed to verify that [redacted] was effective over the desired range of ability levels ($-3 < \theta < 0$).

Findings
The comprehensive evaluation and review process described above has allowed for the following findings: 1) <u>Evaluation of Test Items</u>: When looking at the item development and review processes that were followed with [redacted], the policies and procedures that were followed are consistent with expected practices as described in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME), the *European Test User Standards* (EFPA, EAWOP), the *European Test Review Model* (EFPA, EAWOP), and other key sources that define best practices in the testing industry. Specifically, the test items were determined to be error free, unbiased, and were written to support research-based instructional methodology, use culturally-sensitive language and appropriate content-based vocabulary, and assess the applicable content standard. 2) <u>Field Testing</u>: Following a review of the field-testing rationale, procedure, and results for [redacted] the methods and procedures that were followed are generally consistent with expected practices as described in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME), the *European Test User Standards* (EFPA, EAWOP), the *European Test Review Model* (EFPA, EAWOP), and other key sources that define best practices in the testing industry. Specifically, the

field-testing design, process, procedures, and results support an assertion that the sample size was sufficient and that the item-level data were adequate to support test construction, scoring, and reporting for the purposes of these assessments.

3) <u>Evaluation of Test Administration:</u> Following a review of the test administration policies, procedures, instructions, implementation, and results for [redacted], the intended policies and procedures that were followed are generally consistent with expected practices as described in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME), the *European Test User Standards* (EFPA, EAWOP), the *European Test Review Model* (EFPA, EAWOP), and other key sources that define best practices in the testing industry. Specifically, all aspects of the test administration that were reviewed, such as the item-version random sampling and distribution method, the instructions provided to examinees, and the assessment delivery methods, were consistent with other comparable programs. In addition, reasonable accommodations for applicable disabilities were available upon request when feasible.

4) <u>Evaluation of Scaling and Scoring:</u> Following a review of the scaling and scoring procedures and methods for [redacted] and based on the evidence available at the time of this evaluation, the policies, procedures, and methods are generally consistent with expected practices as described in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME), the *European Test User Standards* (EFPA, EAWOP), the *European Test Review Model* (EFPA, EAWOP), and other key sources that define best practices in the testing industry. Specifically, the measurement model, scoring method, and cut-off score were largely considered to be appropriate and in alignment with comparable industry standards, particularly as it pertains to certification-based proficiency exams. Minor changes were related to differential scoring procedures based on item difficulty were recommended.

5) <u>Evaluation of Psychometric Validity</u>: Following a review of evidence for specific psychometric validity questions for the [redacted], the policies, methods, procedures, and results that were followed are generally consistent with expected practices as described in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME), the *European Test User Standards* (EFPA, EAWOP), the *European Test Review Model* (EFPA, EAWOP), and other key sources that define best practices in the testing industry. Assumptions regarding unidimensionality of the underlying latent trait (Python programming proficiency) were found to be appropriate, as were the item difficulty and discrimination levels. Further, analyses conducted using all available demographic and person-level evidence found no potential sources of bias, differential item functioning, or measurement invariance across demographic groups.

Conclusions Regarding [redacted]

On the basis of all available evidence and the subsequent findings listed above, the following conclusions are deemed appropriate and justifiable:

1) The development and refinement of the [redacted] certification exam has been conducted in a manner consistent with the prescriptive recommendations for best practices presented in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME), the *European Test User Standards* (EFPA, EAWOP), and the *European Test Review Model* (EFPA, EAWOP).

2) The current version of the [redacted] can be considered to be psychometrically valid, reliable, and devoid of test bias in alignment with the guidelines and standards for psychological and educational testing practices put forth by the APA, AERA, NCME, EFPA, and EAWOP.