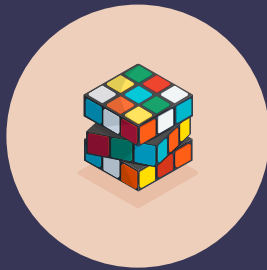
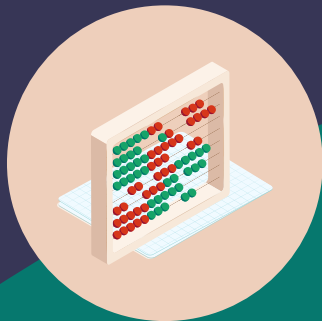


Cognitive Assessments White Paper

In partnership with Psycholatte



Introduction

- Intended audience 3
- Applicability of the tests 3
- Usefulness of the tests 3

The Tests

- Verbal Comprehension..... 5
- Numerical Comprehension..... 6
- Abstract Reasoning..... 7
- Attention and Focus 8

Psychometric Methods

- Item Response Theory (IRT) 10
- Computerized Adaptive Testing (CAT) IRT 11

Item Bank Construction

- Stage 1: Item creation and generation.....13
- Stage 2: Piloting.....14
- Stage 3: Parameter estimation 15
- Stage 4: Online parameter estimation..... 16

Item Bank

- Scoring 20
- Next item selection 20
- Stop criteria..... 21

Test Results

- Observed scores24

Discussion

- Test security.....27
- Test fairness.....27
- Reliability27
- Validity.....28
- Test fairness.....28

References29

Introduction

Intended audience

This is a technical document. It contains data and graphs, references to psychometric and statistical methods and provides detailed information about the quality of the Workable Assessments.

We have tried to make this document as accessible as possible. A trained psychometrician should be able to get the information they need in order to make an informed decision about the quality and appropriateness of the tests for a specific personnel selection scenario.

Applicability of the tests

The tests were designed for the exclusive use of Workable. Workable provides a platform for the assessment of candidates for all kinds of jobs in almost all geographies. After running an analysis of the job postings in the Workable database, we found that the majority of positions posted by Workable users were for white collar jobs (between 70 and 80%). In order to measure this, we used Workable's proprietary AI classification system to classify each job and manually reviewed a randomly selected sample of 80 job postings.

Based on the job classification analysis, we assumed that the test-taking audience would be college graduates. Since college education is diverse and varies on the subject matter of study, we assumed that the common denominator of 12 years of education is a good predictor of people's basic knowledge and comprehension. Comparison of different curricula across countries and generations (curricula change over time) was not performed during the development of this project.

Usefulness of the tests

The four cognitive tests were modeled after the most common cognitive assessments in the testing industry.

The abilities measured have been shown to predict future job performance (predictive validity) for a variety of job positions (Borman et. al. 1997; Hough & Oswald 2000; Ryan & Ployhart 2014; Sackett & Lievens 2008; Salgado 2017). Nevertheless, the fitness of each cognitive measure for a specific position has to be assessed on a case by case basis.



The Tests

Verbal Comprehension

Communicating in written form is a universal requirement for white collar jobs. From reading simple instructions to understanding directives and legal documents, employees are expected to comprehend written material. Sometimes, verbal comprehension expands into understanding deeper meanings, like spotting areas of uncertainty in a contract or gray areas in an argument.

In our version of verbal comprehension, a passage is presented to candidates and they are expected to find the correct answer.



Food selection & consumption

Tons of edible food is discarded each year. Often, people misjudge the suitability of products for consumption which leads to this. Expiration dates can be misleading, since they are not determined by exact procedures. Experts suggest that a standardization of the date labels might keep consumers from needlessly throwing away edible food. Improvement of packaging might also help keep food fresh for longer periods of time.

Consumers also exhibit cognitive bias. People won't buy vegetables that are the last ones left on the shelf. We tend to think that other consumers have purposely avoided those particular items. Although their reasoning is not clear, we are still reluctant to buy them. This process is coined as "social comparison" by psychologists.

Producers and distributors impose their own standards for discarding food as well. For example, although the shape of a fruit may be orthogonal to its quality, distributors will still arbitrarily determine the acceptable shape of a particular fruit. As a result, consumers are gradually accustomed to specific visual attributes of vegetables, and adjust their own selection criteria.

Which of the following is the most appropriate summary of photograph 2?

- The last vegetables left on a shelf are probably defective. The number of consumers that have examined and rejected these items is large enough to consider them justly evaluated through a process of social selection.
- People choose their produce rationally, and items left last are probably defective somehow. By a process of social comparison, consumers are justified to reject food left back at the shelf at the end of the day.
- People are prone to erroneously perceive an item that is the last one on the shelf has been actively rejected by other consumers. Because we are accustomed to take other people's behavior into account when making our own decisions, we tend to be reluctant to buy those products, but we do so in error.
- Some people tend to be deliberate and compare their produce against a set of established criteria, whereas others use superficial criteria such as the food's appearance, in deciding whether to buy an item or not.

Numerical Comprehension

It's not uncommon for an employee to have to understand relationships between magnitudes, solve mathematical problems or generally process reality in a mathematical way. We opted to focus on the one ability that the advent of spreadsheets is not able to make up for: The creation of mathematical models from real life situations.

An on-screen calculator is provided to the candidate. Numbers are highlighted and when clicked, they will be entered correctly into the calculator. Therefore, the candidate can focus only on finding the exact formula that solves the problem.



For a given product, in 2011, 112,965 items were sold for a total value of 180,744. How much does a unit of this product cost?

Please use a point (.) as your decimal separator



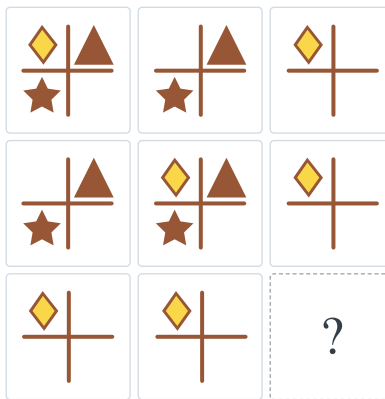
You can use your keyboard for faster input!

AC				
7	8	9	÷	
4	5	6	×	
1	2	3	-	
0	,	=	+	

Abstract Reasoning

There are many cases where an employee will have to identify patterns that are not immediately apparent, or situations where a candidate will have to make inferences based on observations. For this reason, abstract reasoning is commonly used in job roles that require relevant abilities. The test format was originally developed by Dr. John C. Raven in 1938. Such tests measure a facet of general intelligence.

In this abstract reasoning test version, one of nine elements is missing. The candidate's task is to find the missing element among several suggested elements using the multiple choice paradigm.



What replaces the question mark?



A



B



C



D



E

Attention and Focus


Testing for attention to detail is usual within the assessment industry. This skill was useful in the previous decades when data entry was an important part of many occupations.

In recent years, human - computer interaction time has increased. Barcodes, QR codes, NFC, online newspapers, government public databases, online maps, all ensure that data is not entered by humans when aspects of the real world need to be input to a computer.

This is why we opted to focus on the one thing that is much sought after in white collar occupations: Attention itself.

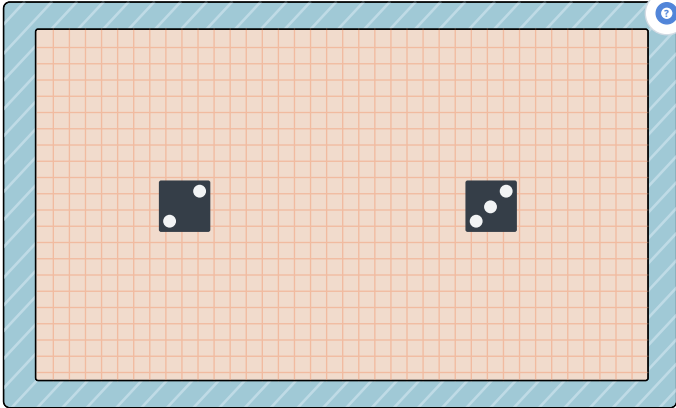
In the always-connected world of social media and smartphones, it has become evident that the attention span of each generation is constantly diminishing (Subramanian 2018). Being able to deeply focus might be one of the highly sought abilities going forward.

In our 'Attention and Focus' test, the candidate has to start from the center and according to the coloured pattern, apply simple mathematical functions to the numbers as presented in dice form. If the intermediate result is forgotten, this suggests attentional disruption, since it's very hard to remember where one has stopped.





Guide to calculate dice results


- the enclosed dice number +1
- the enclosed dice number -2
- the largest dice number +1
- the smallest dice number -1
- the enclosed dice number -1
- the largest dice number +2
- the smallest dice number +1
- the largest dice number -1





What is the final result of the box above?



A



B



C



D


E


F


G


H


I



Psychometric Methods

Item Response Theory (IRT)

Item Response Theory (IRT) for our purposes can be described as a questionnaire/test scoring method. This means an ability test, such as an exam with multiple choice answers where one is correct, or a trait test, such as a personality test with likert type options. This document refers to ability tests.

Some of the IRT's advantages are: (1) it estimates several properties of individual questions or items, and (2) it takes into account these properties in scoring. An example of such a property is the item difficulty, so by knowing how difficult an item is, we can take that into account in scoring. An oversimplified example is that more difficult items carry higher scoring marks. In contrast, in classical scoring methods, each question carries a point and so, classical scoring techniques do not differentiate items in scoring.

An IRT implementation provides empirical item estimates using statistical methods. The statistical values describe the items in terms of difficulty, discriminability and other attributes and, so, the issue of subjectivity in interpreting the quality and the characteristics of items is minimized. In addition, this allows us to describe the items' or tests' appropriateness for different populations. For instance, particularly difficult questions are better suited for populations of high ability, and particularly easy questions are better suited for populations of low ability.

Nonetheless, the main advantage of IRT is that it can provide scoring with incomplete responses. That means that there is no need to answer all questions to estimate a score for a participant.

Computerized Adaptive Testing (CAT) IRT

Imagine a classroom where a teacher examines a student during a lesson by asking a series of questions. They do not have time to ask 20 questions. They ask an average difficulty question. The student responds. In their mind, the teacher assesses the response and creates an early understanding of the student's knowledge. Based on that understanding, they will follow up with an easier or harder question. They will keep asking appropriate questions until they feel confident enough that their understanding of the student's knowledge is correct. What if we could program a computer to do that?

IRT Computer Assisted Testing is the method of using a computer to score and shape a testing experience unique to a candidate. In real time, after each question, the computer will score the candidate and assess the statistical confidence of the score. If this statistical confidence (Standard Error of Measurement) is satisfactory, it will stop the test. If not, it will select the next most appropriate question from a pool of available questions usually called an "item bank".

This has many advantages:

- The test is shorter (fewer questions) and takes less time.
- Not all items have to be presented to the candidate.
- The items presented are dependent on a candidate's performance.
- The test ends when the algorithm is certain enough about a candidate's ability.



Item Bank Construction

Stage 1

Item creation and generation

For the verbal and numerical tests, items were created by psychologists with considerable experience in item creation and psychometrics. The created items were then reviewed by two more psychologists with similar credentials.

For the abstract reasoning and attention and focus tests, about 3-4 prototypes were created and considered. After consideration, reviews and small-scale tests, the final designs were programmed into test generators that were able to produce many test items. The test generators' output was reviewed many times in order to make sure all conditions and limits were met and more conditions were decided if needed. Finally, the abstract test's output was also reviewed by humans in order to make sure that when shapes overlapped, no shape was hidden (e.g. a dotted circle printed over a solid circle of the same size).

Stage 2

Piloting

For each test, 50 items were created and then split in two batches of 25. The two batches were as similar as possible. Wherever the items were generated by algorithm, the two batches were created with equal parameters.

Workable employees were randomly assigned to complete one of the two batches. Each participant answered every question in their assigned batch. A person that completed all four tests, would have completed 100 items in four sittings.

The data was analyzed and questions were removed if they didn't contribute to the reliability of the test, or were too easy, too difficult, or ambiguous.

For the verbal test, 35 questions from the initial 50 were kept. Since some questions were dependent on passages, we decided that some passages had to be removed altogether. The Cronbach α achieved was .82 and .85 for batches 1 and 2 respectively.

For the numerical test, 47 questions from the initial 50 were kept. The Cronbach A achieved was .83 and .88 for batches 1 and 2 respectively.

For the abstract test, 47 questions from the initial 50 were kept. The Cronbach α achieved was .87 and .81 for batches 1 and 2 respectively.

For the attention test, 50 questions from the initial 50 were kept. The Cronbach A achieved was .89 for both batches. We ran the pilot process twice because the previous test structure had a format that meant if the candidate took notes, the test would become very easy.

Stage 3

Parameter estimation

Parameter estimation was performed using Method A (Baker & Kim 2004), from the open source irtplay package (Hwanggyu 2020). The scaling constant was set to $D=1$, the model was a one parameter model with the discriminator parameter set to be free and same for all items. Neither prior distributions, nor prior values for the parameter estimation were used.

Before implementing a CAT IRT questionnaire we need an item bank, that is a set of items with specific parameters. These parameters are used for score estimation and in the choice of the items each participant is exposed. The pilot study served to estimate these parameters.

For the abstract questionnaire, we used a sample of 79 participants to estimate parameters for 47 items. For the attention questionnaire, we used a sample of 82 participants to estimate parameters for 50 items. For the numeric questionnaire, we used a sample of 86 participants to estimate parameters of 47 items. Finally, for the verbal questionnaire we used a sample of 81 participants to estimate parameters for 35 items.

The items used to perform these tasks are generally referred to as test items. After estimating the parameters of these items it is possible to perform CAT IRT scoring.

Stage 4

Online parameter estimation

In this implementation, we needed a larger item bank and therefore additional items were required while the CAT was running. We refer to the items added in that stage as trial items.

The additional steps we followed have been summarized below:

1. Distributed trial items randomly and collected responses for these items. In this stage, new items did not have parameters and so were not used for scoring candidate results.
2. Evaluated trial items using the empirical responses.
3. Analyzed trial item responses in order to estimate IRT parameters.

After parameters were estimated and found acceptable, they were added to the item bank. CAT scoring was then performed using more test items and new items were added for trialing.

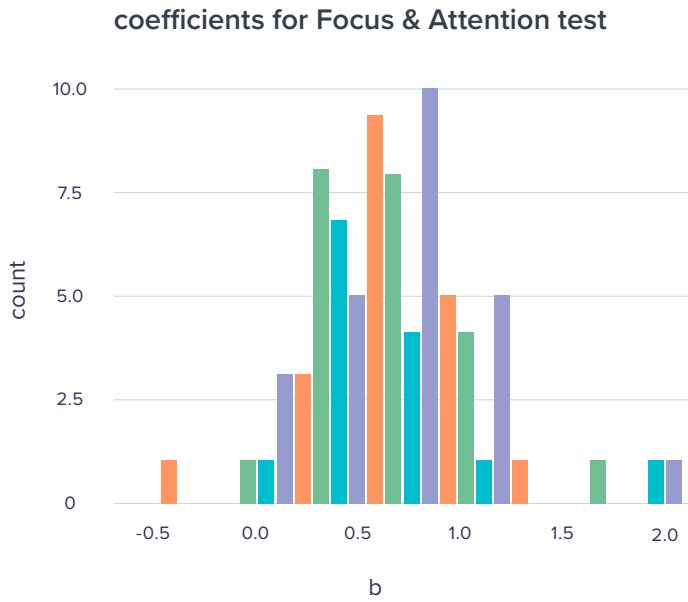


Item Bank

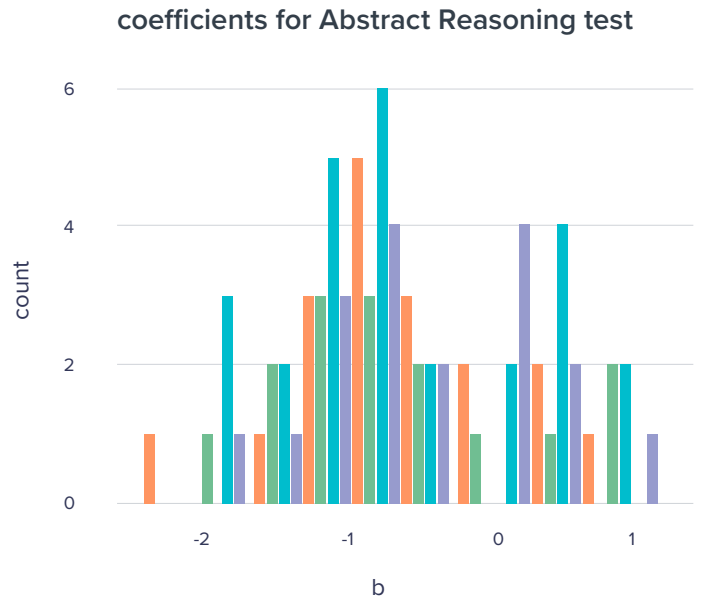
In its current form, the abstract, attention, numerical and verbal item banks have 75, 79, 93 and 72 estimated item parameters respectively.

In a one parameter model (1PL) the estimated parameter b represents each item's difficulty. Therefore, a high b coefficient represents a difficult item, and a low b coefficient represents an easy item. In a 1PL model, the discrimination parameter a is constant and same for all items. The table below shows descriptive statistics for the parameters in each questionnaire.

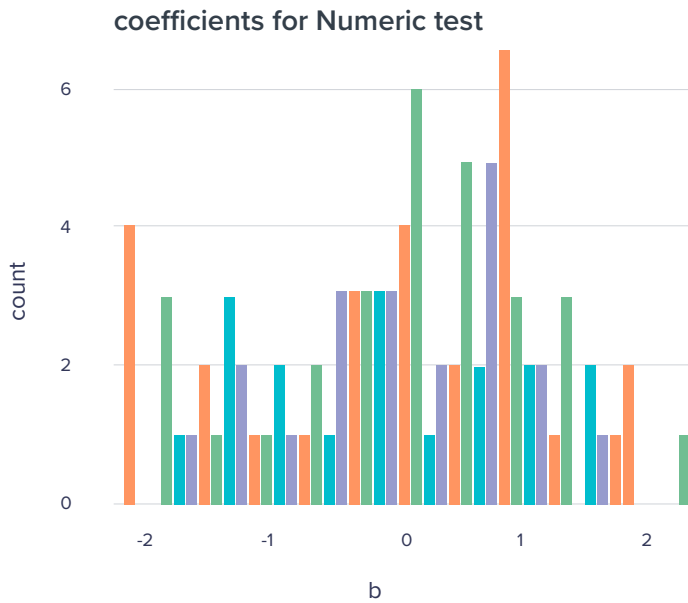
Test	A				
	Mean	SD	Min	Max	
Abstract	1.95	-0.68	0.82	-2.54	1.08
Attention	2.01	0.68	0.39	-0.39	2.05
Numerical	2.05	-0.02	1	-2.11	2.32
Verbal	1.9	-0.98	0.72	-2.68	0.73



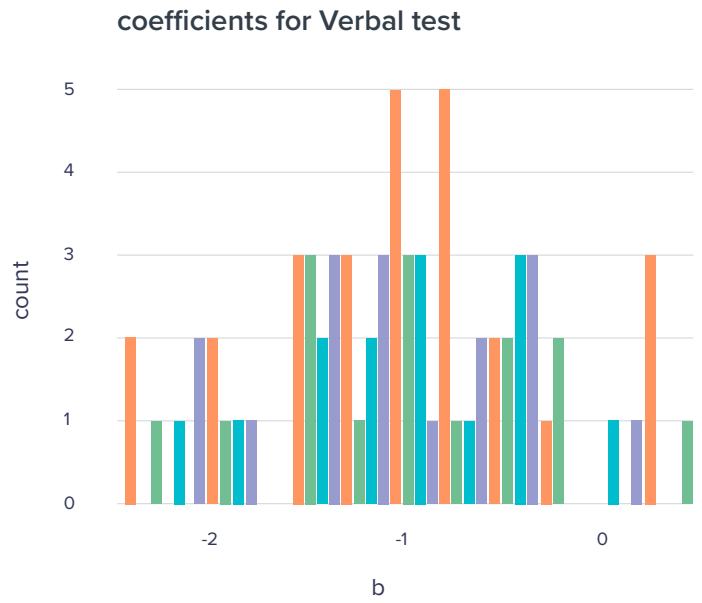
Observations = 79
 Mean = 0.68
 SD = 0.4
 Median = 0.64



Observations = 75
 Mean = 0.68
 SD = 0.82
 Median = 0.88



Observations = 93
 Mean = 0.02
 SD = 1
 Median = 0.06



Observations = 72
 Mean = 0.98
 SD = 0.72
 Median = 1.01

Figure. Histogram of beta coefficients

The test information function (TIF) is another relevant statistic. The test information function indicates the theta values a test is able to estimate accurately.

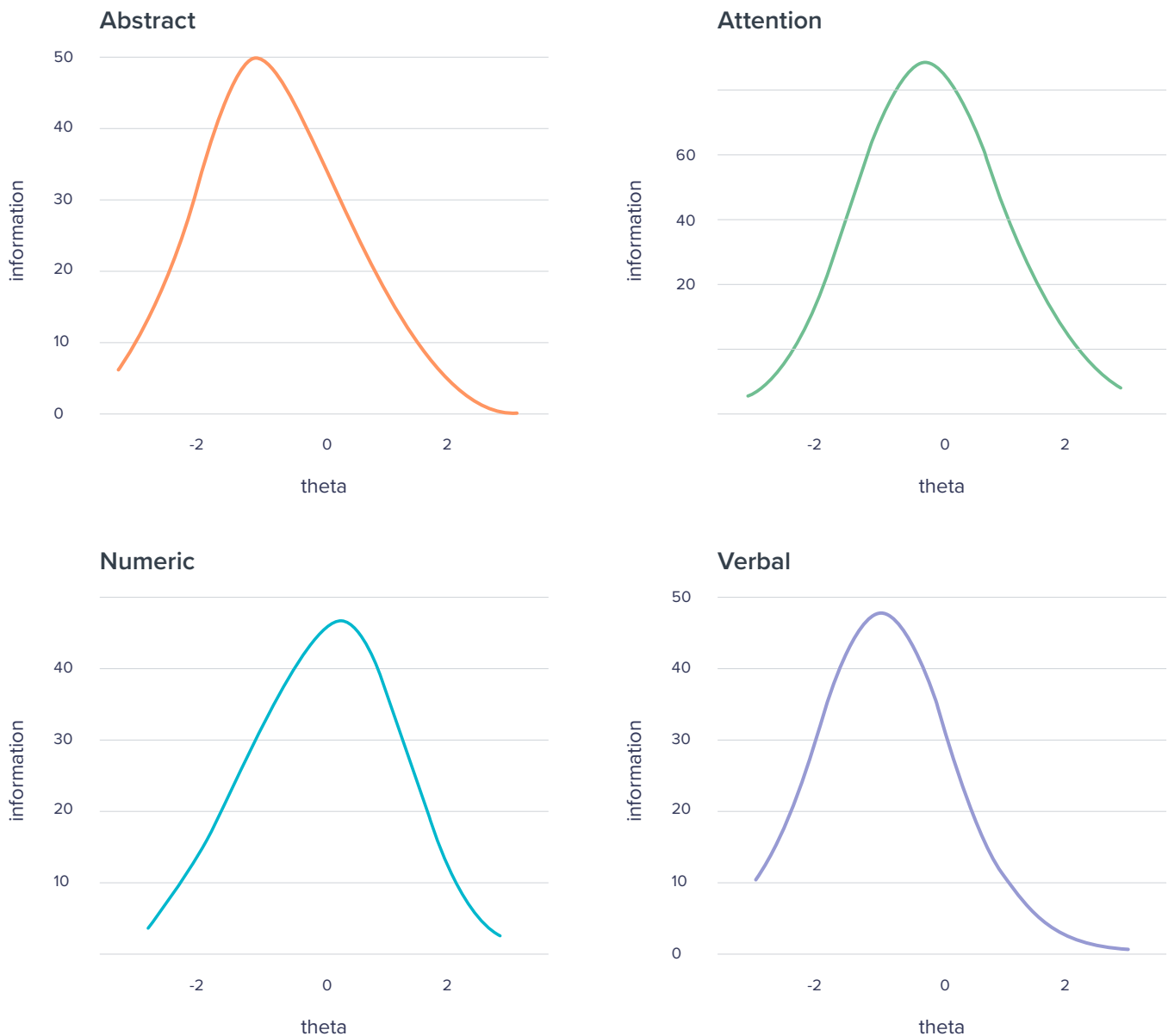


Figure. Test information Functions for the four tests

Scoring

Responses are scored using the Expected A Posteriori (EAP) estimation method. The EAP parameters in this implementation is a normal prior distribution ($\mu=0$, $\sigma=1$), with theta limits -3, 3, and 60 quadratures.

Next item selection

At the beginning, three items are asked which are selected with assumed theta scores of -1, 0 and 1. Then the theta value of the candidate is estimated and the next item is selected. This process is repeated until a stop criterion is reached.

During the course of a questionnaire, theta scores are estimated in real time after each response. The current theta score is used to choose the next item. The next item is chosen randomly among the five items that maximize information for the currently estimated theta score. The selection process ensures that the same item is not shown twice in the same assessment.

The algorithm is much more complicated for the verbal comprehension test. This test contains passages and subsequent comprehension questions for each passage. So, after reading a long passage, the candidate is expected to answer more than one question. For each passage at least three questions are asked. Still the algorithm tries to present the most appropriate question for the estimated theta value of the candidate.

Stop criteria

The main stop criterion is that the standard error falls below 0.55, or less than 20 questions are answered. This value is roughly equivalent to a Cronbach A score of 0.70. Another stop criterion is when 20 items have been presented to the participant.



Test Results

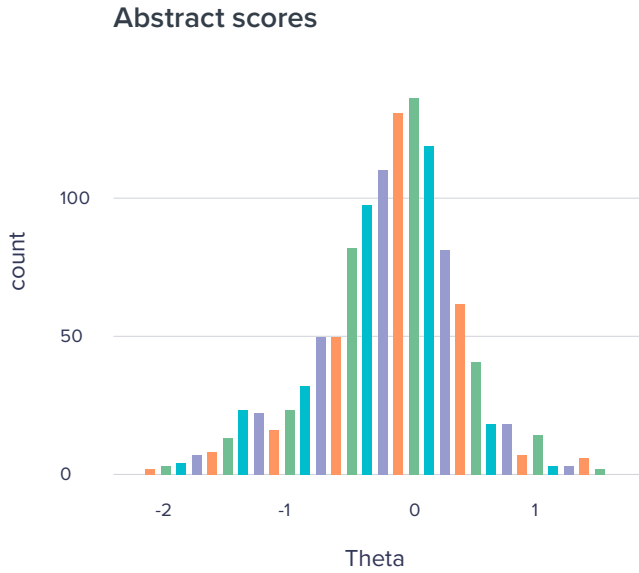
The figure below indicates how many questions participants answer before the test stops. In general, for the abstract and the numerical tests, the majority of participants answer around 11 questions. In comparison, the verbal ability test data is more dispersed and participants normally answer more questions before the test ends.



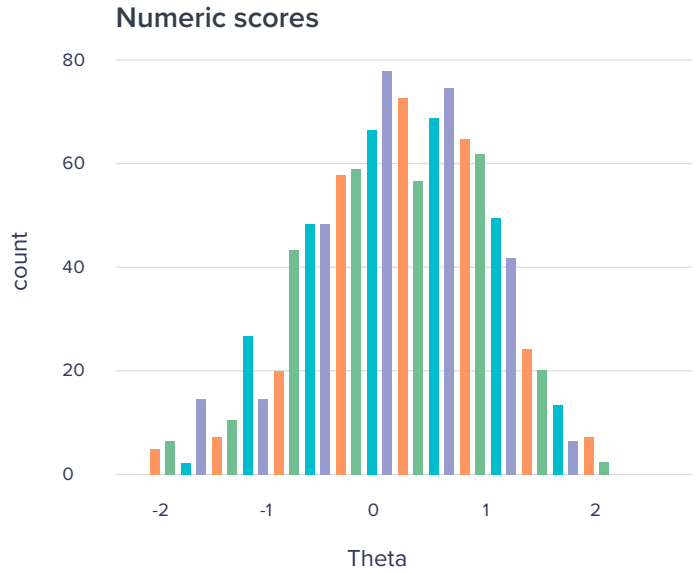
Figure. Number of items answered per participant

Observed scores

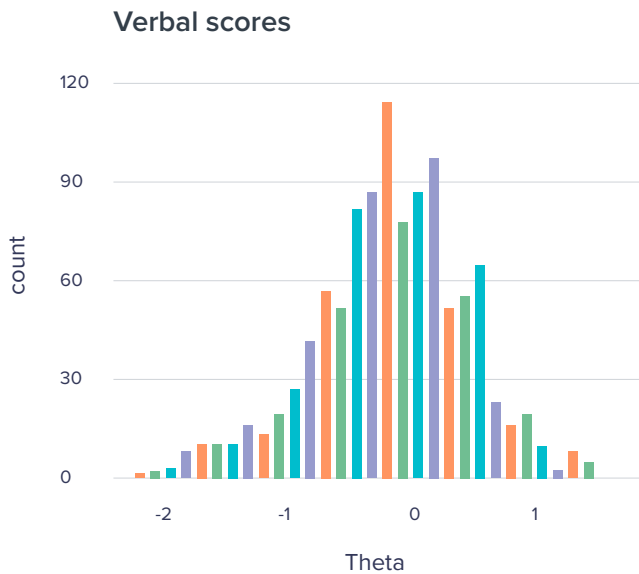
The distribution of the final scores is shown in the figure below.



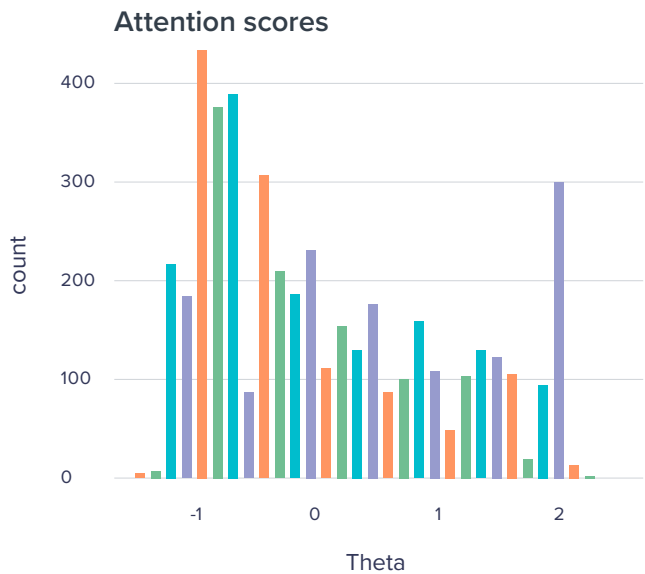
Observations = 1139
 Mean = 0.08
 SD = 0.58
 Median = 0.03



Observations = 1043
 Mean = 0.32
 SD = 0.8
 Median = 0.36



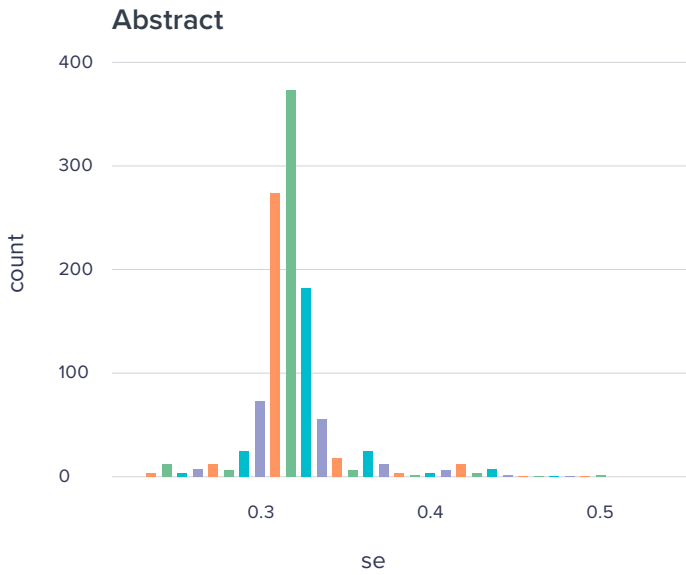
Observations = 1057
 Mean = 0.07
 SD = 0.61
 Median = 0.07



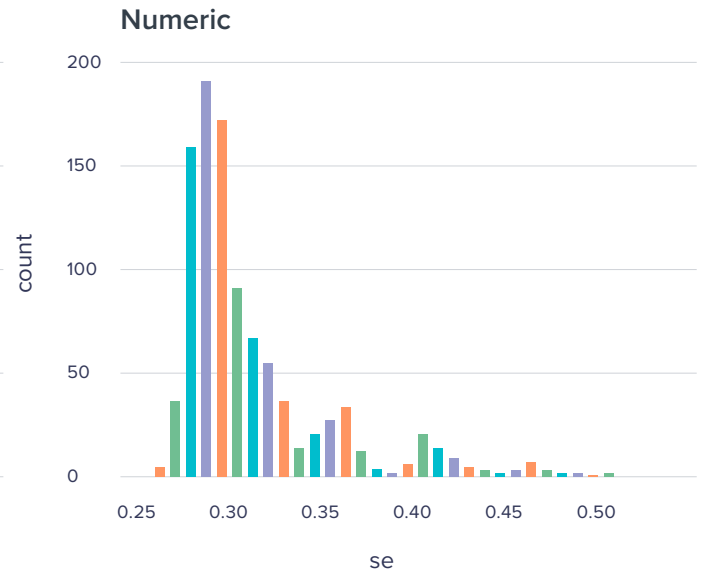
Observations = 4672
 Mean = 0
 SD = 1
 Median = 0.25

Figure. Histogram of theta scores

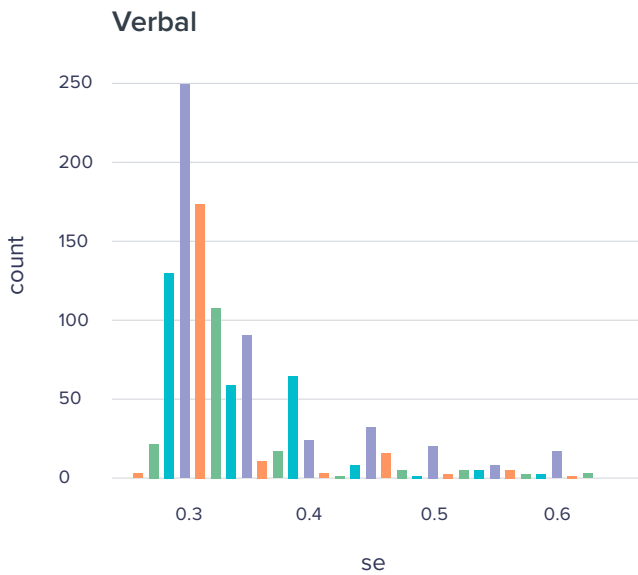
The distribution of the standard error of the scores is shown below. As mentioned, the test stops when the standard error (SE) is below 0.55, however for the abstract and numerical tests, the last question achieves a lower SE. The reason some observations of the verbal test are a bit higher than .55 is because the other stop criterion is 20 questions.



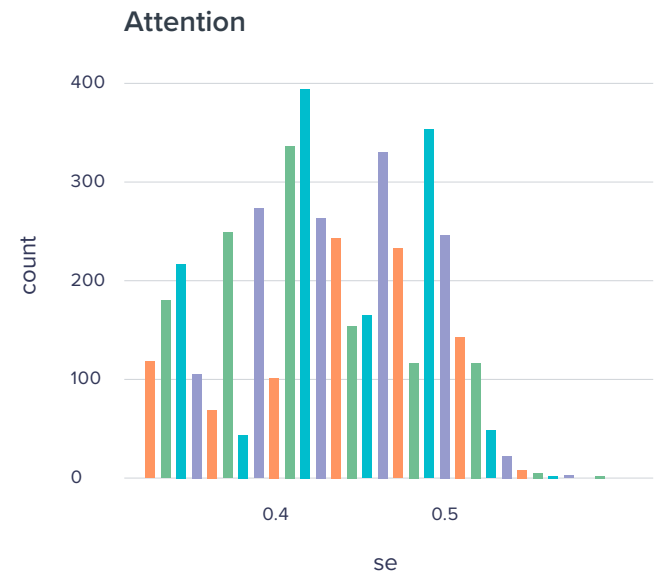
Observations = 1139
 Mean = 0.32
 SD = 0.03
 Median = 0.31



Observations = 1043
 Mean = 0.31
 SD = 0.04
 Median = 0.29



Observations = 1057
 Mean = 0.33
 SD = 0.06
 Median = 0.31



Observations = 4672
 Mean = 0.43
 SD = 0.05
 Median = 0.43

Figure. Histogram of standard error of the scores



Discussion

We present four adaptive, IRT-based ability tests. Based on the data and procedures presented above, we summarize the main points defining the quality of these tests.

Test security

The test is adaptive, which means that the test content presented depends on the performance of each candidate.

It is not possible to create a stable, predictable path through the test. Each time, the algorithm selects the five best items for the current estimate of the candidate's ability. It then picks one at random. This means there are roughly 5^{10} different paths for a candidate presented with 10 questions.

Furthermore, the test content travels from Workable's server to the candidate's browser shortly before presentation. All communication between the server and the browser is encrypted and Workable employs data protection audited by standard ISO27001 procedures.

Test fairness

We are recording gender, ethnicity and age, in order to measure and correct for any possible bias. At this time our sample is not representative of enough countries and occupations to make any statistical inferences.

Reliability

The tests present a very strong reliability. Although the algorithm is set to terminate the test at Standard Error of Measurement 0.55 or less, most of the average SEM scores are between 0.31 to 0.33. This happens because there is a minimum number of questions asked. When converted to Cronbach A, the tests present a reliability of about 0.9.

Test	SE Meas	Cronbach's A
Verbal	0.33	.891
Numerical	0.31	.904
Abstract	0.32	.898
Attention	0.42	.815

Validity

For the use of tests in the hiring process, the ideal approach to validity is the predictive validity. That is, the test performance should be correlated with the job performance one year later. Many studies have already established such correlations with tests similar to ours for a variety of job roles (Schmidt & Hunter 1998; Schmidt et al. 2016; Ryan & Ployhart 2014).

This is the reason why Verbal Comprehension and Numerical Comprehension, along with Abstract Reasoning, are so common in pre-selection testing.

The candidate sample we have today is not representative of all geographies Workable operates in, or of all the economy sectors, or of all age groups. When we collect more data we will be able to support test validity further by comparing test performance across different economic sectors.

Test fairness

It is too early to conclude anything about test fairness. The sample we currently have is mostly based on one, non-technical, highly educated job profile. The differences we are noticing are small (0.3 Standard deviations). Test fairness will need to be reviewed when the test is exposed to more countries and more sectors of the economy.

References

- Baker, F. B., & Kim, S. H. (2004). Item response theory: Parameter estimation techniques. CRC Press.
- Borman, W. C., Hason, M. A., & Hedge, J. W (1997). Personnel Selection. Annual review of psychology, 48, 299-337. <https://doi.org/10.1146/annurev.psych.48.1.299>
- Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future-- remembering the past. Annual review of psychology, 51, 631-664. <https://doi.org/10.1146/annurev.psych.51.1.631>
- Hwanggyu Lim (2020). irtplay: Unidimensional Item Response Theory Modeling. R package version 1.6.2. <https://CRAN.R-project.org/package=irtplay>
- Ryan, A. M., & Ployhart, R. E. (2014). A century of selection. Annual review of psychology, 65, 693-717. <https://doi.org/10.1146/annurev-psych-010213-115134>
- Sackett, P. R. & Lievens, F. (2008). Personnel selection. Annual review of psychology, 59, 419-450. <https://doi.org/10.1146/annurev.psych.48.1.299>
- Salgado, J. F. (2017). Personnel Selection. Oxford Research Encyclopedia of Psychology.
- Salgado, J. F. (2017). Using ability tests in selection. The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection and Employee Retention, 1st ed
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin, 124(2), 262–274. <https://doi.org/10.1037/0033-2909.124.2.262>
- Schmidt, F., Oh, I-S & Shaffer, J (2016). The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 100 Years of Research Findings. https://www.researchgate.net/publication/309203898_The_VValidity_and_Utility_of_Selection_Methods_in_Personnel_Psychology_Practical_and_Theoretical_Implications_of_100_Years_of_Research_Findings
- Subramanian K. R. (2018). Myth and Mystery of Shrinking Attention Span. International Journal of Trend in Research and Development, 5(3), 1-6.



Visit **workable.com** to discover all the ways Workable helps
you **find, evaluate** and **hire** the **best candidates**.

