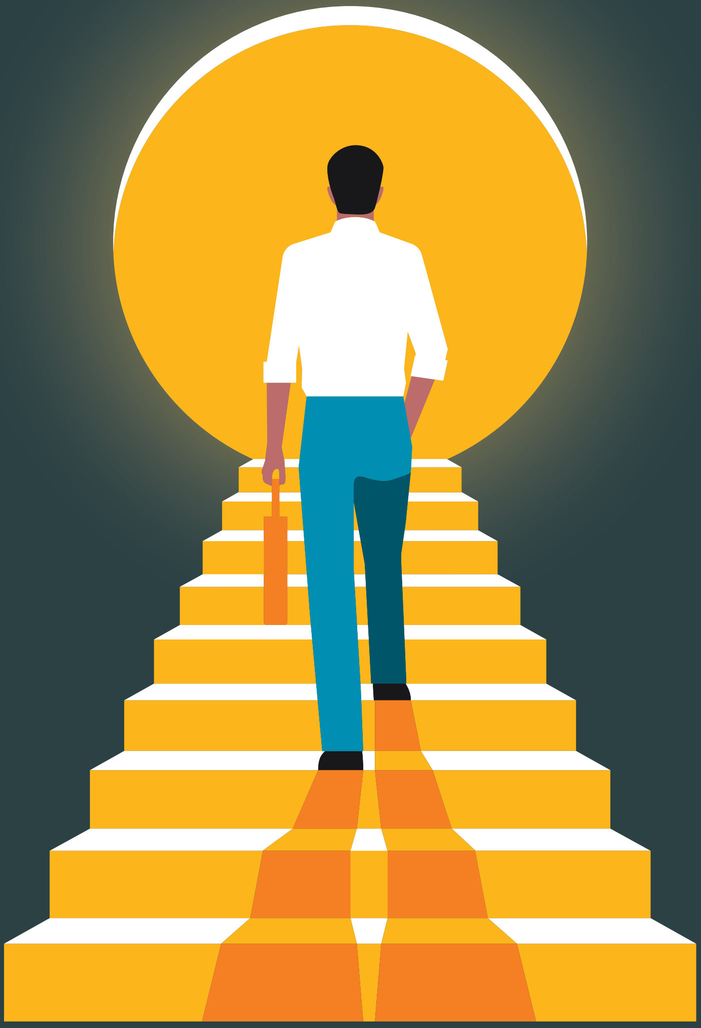
Pearson

GED Testing Service

Efficacy Research Report

April 2018



Contents

03 Introduction

05 Product summary

07 Assessment quality indicator

08 Foundational research

09 Product research

Introduction

In 2013, Pearson made a commitment to efficacy: to identify the outcomes that matter most to students and educators, and to have a greater impact on improving those outcomes. Our aspiration was to put the learner at the heart of the Pearson strategy; our goal was to help more learners, learn more.

A critical part of Pearson's portfolio is its Assessment business, which is really a services business supporting our customer requests by designing, building, administering, scoring, and reporting on test-taker performance in many different contexts (ranging from K–12 classrooms to the workplace) and for different purposes (ranging from supporting classroom instruction through ongoing progress monitoring to certifying fitness for employment in a given occupation). The people who take these tests are learners on a journey, similar to students who use our courseware products in the classroom to fulfill course requirements. In this case, however, the test is serving a slightly different function along this journey than would one of our digital courseware products. Taking a test is not a learning experience in and of itself, but rather, the scores and diagnostic information from these assessments may be used by instructors and others to make decisions about a learner's progress along their journey. Therefore, a measure of efficacy for assessments is not whether taking the test leads directly to higher achievement or passing the course, but whether the scores and other diagnostic information provide an accurate snapshot of what the learner knows and can do. In other words, the efficacy of an assessment is its fitness for a given purpose.

The fitness of an assessment for a given purpose, in turn, is defined by three primary qualities or attributes of test scores and their use: validity, reliability, and fairness. The *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) have defined these attributes as follows:

— **Validity** is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). Validity requires evidence that test scores can be interpreted as they are intended and can be appropriately used for a specific, defined purpose.
— **Reliability** is "the consistency of scores across replications of a testing procedure" (p. 33). Reliability requires evidence of the consistency of scores over time, across multiple forms of the assessment, and/or over multiple scorers.
— **Fairness** suggests that "scores have the same meaning for all individuals in the intended population" (p. 50). Fairness requires evidence that when assessments are administered as intended, items are not systematically biased against any particular group of test-takers and students are not hindered in demonstrating their skills by irrelevant barriers in the test administration procedures.

Given the longstanding role of the Joint Committee Standards as a source of guidance on best practices in the development and evaluation of tests and the role these standards play in the legal defensibility of assessment, Pearson has adopted these three attributes as the Assessment Quality Indicators on which we publicly report evidence underlying our assessment products. Each attribute is associated with a range of evidence types that are more or less relevant in a given context depending on the test's particular purpose and intended uses. For example, there are five commonly-accepted types of validity evidence that can be woven together to formulate an argument that a particular test can be interpreted as intended and used in a particular way, including evidence about how the assessment content was developed and how scores on the assessment relate to scores on other measures of the same kinds of knowledge and skills (AERA, APA, NCME, 2014). Similarly, there are different indices of reliability that can be provided, depending on the purpose and implementation of the test — when and how often it is administered, how it is scored, and how scores are reported. Such indices might include the average inter-item correlation or correlations between scores from different forms of the assessment, or across different times when the assessment is administered. Finally, fairness can also be supported by different types of evidence, including the results of analyses that specifically attempt to isolate items that appear to function differently for people in different subgroups (e.g., males versus females) and results from analyses of item content by specially formulated expert committees whose purpose is to identify potentially biasing content.

Pearson's assessment products are designed, built, and maintained over time by teams of subject matter experts and Ph.D. level research scientists trained in the science of assessment. These teams regularly (in some cases, annually) carry out studies to collect the kinds of validity, reliability, and fairness evidence described above, in accordance with the Joint Committee Standards. This evidence is typically consolidated and published in a technical manual or technical report that is updated with each new revision of the test. For that reason, much of the research we summarize on our assessment products has been completed internally and in many cases, we refer the interested reader to the technical manuals for full details of the research studies and associated evidence.

Special thanks

We want to thank all the customers, test takers, research institutions and organizations we have collaborated with to date. If you are interested in partnering with us on future efficacy research, have feedback or suggestions for how we can improve, or want to discuss your approach to using or researching our assessments, we would love to hear from you at efficacy@pearson.com.

Kate Edwards

Kate Edwards
Senior Vice President,
Efficacy and Research, Pearson
April 3 2018

Product summary

The GED test has been in existence since 1942 when its primary purpose was to facilitate US veterans returning from World War II obtaining a high school credential and entering college or joining the workforce. The test was revised three times between the 1970s and the early 2000s. In 2014, in response to an increasingly competitive labor market and a growing skills gap in the workforce, the GED test underwent a significant fourth revision, and a complementary GED Program of supporting tools and services was launched. The new test and program is designed to help adult learners obtain the credentials needed to access a variety of postsecondary education pathways and to progress in the workplace and in job training programs.

The revised version of the test allows learners who score at the highest level to achieve college credits, which could further encourage them to pursue post secondary education. Another significant component of this revision to the GED test was the transition to a digital platform. The GED Program is founded on two key assessments with related but differing purposes: the flagship GED test, the nationally recognized high school equivalency test developed by experts with a 75 year history in high school equivalency; and the GED Ready Official Practice Test (GED Ready), which provides a score predictive of performance on the operational GED test.

Both tests were developed concurrently, with similar specifications, yet are used differently. GED Ready was built to the same content specifications as the GED test, and therefore allows individuals to evaluate their likelihood of success on the GED test. GED Ready also provides an opportunity to gain additional testing experience, as well as exposure to the content and format of the GED test.

The GED test itself is intended for jurisdictions to award a high school–level credential. The GED test is also an indication of readiness for some jobs, job training programs, and/or credit-bearing postsecondary education coursework. The GED test:

- Aligns with current high school standards (including US grade 12 standards) and career and college–readiness expectations
- Uses primarily computer-based delivery, which provides a consistent testing experience worldwide
- Has same-day scoring on all four subject area parts and includes an enhanced score report for targeting instruction and remediation
- Offers three score levels:
  - GED Passing Score — at or higher than the minimum needed to demonstrate high school equivalency–level skills and abilities in the US
  - GED College Ready — at or higher than the minimum needed to demonstrate readiness for credit bearing college courses in US
  - GED College Ready+Credit — at or higher than the minimum needed to qualify for granting of up to 10 semester hours of college credit
- Provides an opt–in feature enabling learners to receive text messages sent from the program to help support them and keep them on track with their progress towards attaining their GED credential
- Includes the GEDWorks initiative, in which GED Testing Service is partnering with major employers to provide instruction, coaching, and other support to their employees who are seeking to attain their GED credential

The program operates primarily in the US, although there is also an international presence.

Some of the key components of the GED Program include the following:

- MyGED: the portal through which learners access all of the resources and processes that facilitate their test-taking journey, including accessing tutorials, connecting with preparation programs and materials, viewing test results and score reports, taking practice tests, and career exploration.
- GED Ready: the Official Practice Test that gives learners a prediction of their score on the operational GED test.
- Enhanced Score Report: provides learners with their score, performance level, and descriptions of the skills they demonstrated on the test and those they need to develop further to obtain a higher score.
- New GED test: the new assessment is aligned with current college and career readiness standards.
- Pathsource: a free tool for learners to use to explore a variety of career pathways.
- GEDWorks: an employer–sponsored program supporting employees in the achievement of their GED credential, described in more detail below.
- GED College Ready and GED College Ready+Credit: updated test performance levels enabling high performing students to bypass remedial courses, avoid placement testings and even earn college credits depending on their level of GED test performance.

Additional enhancements will continue to be implemented on an ongoing basis, all of which are designed to improve student outcomes and increase efficacy.

While users tout the ease and convenience of the MyGED portal, and more than 770,000 learners have online accounts, GED Testing Service is currently conducting further research with the goal of continuing to improve overall engagement on, and effectiveness of, the platform. The number of MyGED users who take the GED Ready (available to students through the MyGED portal) and complete the full battery of tests shows initial evidence of active user engagement on the platform.

In addition, an analysis of 2014 – 2016 results demonstrate that 51% of candidates prepared with the GED Ready Official Practice Test, surpassing the company's target of 33%. This target is based on historical percentages of test takers who completed the Official Practice Test as part of the previous GED test edition. Surpassing this benchmark shows strong improvement over time in the number of test-takers who prepare for the GED test. As the GED test evolves and students use the practice test and other tools as part of their preparation strategy, the program aims to have more learners complete and pass the full GED test at increasing rates over time.

GED Testing Service has made a public commitment to ensuring that the test content and performance levels will be monitored over time and kept current with corresponding content and performance requirements, leading to awarding regular high school diplomas.

Assessment quality Indicators

Ensuring the quality of the interpretation of test scores is vitally important, since passing the GED test is a prerequisite to earning or receiving a credential. Quality can be characterized as a function of the fairness of the assessments (*fairness*), the consistency and accuracy of scores (*reliability*), and the extent to which the assessment allows test users to make sound interpretations of GED test-takers' intellectual functioning (*validity*). This section outlines the Assessment Quality Indicators (AQIs) agreed for the GED Testing Service.

AQIs relating to fairness

Assessment quality indicator 1

Test scores can be interpreted the same way for test–takers of different subgroups.

The GED Testing Service strives to provide scores that can be interpreted in the same way for all test-takers, regardless of demographic characteristics, such as gender or race/ethnicity. Fairness implies that when the assessments are administered as intended, test items are not systematically biased against any particular group of test-takers, and students are not hindered in demonstrating their skills by irrelevant barriers in the test administration procedures.

Assessment quality indicator 2

The GED Testing Service provides reasonable and appropriate accommodations to individuals with documented disabilities who demonstrate a need for accommodations.

While accommodations are not a guarantee of improved performance or test completion, it is a goal of the GED Testing Service to provide candidates with full access to the GED test. It is important that fair accommodations are made, where appropriate, in line with accommodations procedures.

AQIs relating to reliability

Assessment quality indicator 3

An examinee's GED test scores are stable, in that the scores are consistent, both over time and on multiple test administrations.

An important goal of the GED Testing Service is to minimize errors in judgment and decision making by providing scores that are consistent over different testing occasions and test administrations. This AQI is predicted by an examinee's standard error of measurement (SEM). SEM is an estimate of the average amount of error that is associated with scores derived from a test.

AQIs relating to validity

Assessment quality indicator 4

GED test scores can be interpreted as a measure of examinees' comprehension of course content, and thus, general intellectual ability.

A key goal of the GED Testing Service is to enable test users to make sound interpretations about examinee ability, and to support identification or placement decisions by providing measures that accurately capture general intellectual ability, as well as profiles of relative strengths and weaknesses across different aspects or domains of cognitive ability. It is therefore important that the major content of a high school program of study is reflected in test specifications and test development plans. It is also vital that a single construct underlies the responses to the items on a test, in order to validate the single standard score reported by the GED Testing Service.

Foundational research

The overall design and framework of the GED Program is centered around providing measurements and tools documenting and promoting student achievement. More information about the test itself and the underlying documentation and validity framework may be found in the GED test Technical Manual on the GED Testing Service website.

The philosophy underlying the GED test is that there is a foundational core of academic skills and content knowledge that an adult must demonstrate to be certified as being prepared to enter a job, a training program, or an entry–level, credit-bearing postsecondary course. This foundational core of knowledge and skills is defined by career and college-readiness standards (Pimentel, 2013), now adopted in some form by the majority of states.

The three purposes of GED test are to provide the following:

1. A high school–level credential
2. Information about a candidate's academic strengths and weaknesses
3. Evidence of readiness to enter workforce training programs, careers, or postsecondary education

The GED test is future-focused, designed to provide information about candidate readiness that is directly tied to the next steps in a candidate's preparation and training. Many factors impact actual workplace or postsecondary success, such as engagement, teamwork, or creativity. However, the GED test focuses on those foundational career and college academic skills and knowledge that are critical for candidates to be prepared for the next step in their future, whether they seek to enter the workforce or some form of postsecondary education for further education and training.

Product research

The GED Testing Service team regularly carries out studies to collect the kinds of validity, reliability, and fairness evidence described above, in accordance with the Joint Committee Standards (AERA, APA, NCME, 2014). This evidence has been consolidated and published in a technical manual, which is updated with each new revision of the test. For that reason, much of the research we summarize in the following section has been completed internally. We encourage readers who are interested in the full details of the research studies and associated evidence to consult the official technical manual.

Stand–alone field test

Study citation	Not available
Research study contributors	Not available
Type of study	Stand-alone field test (SAFT)
Sample size	Not available
Description of sample	GED Testing Service recruited a large sample of adults to participate in the SAFT. The participants represented various regions across the United States. The field test included multiple waves of administration across several months. Items were assembled into fixed modules (modules were not necessarily representative of an operational form). A spiral plan was implemented that allowed a minimum of 250 responses to each item type. Technology enhanced items and extended response (ER) and short answer (SA) items were overlapped across multiple modules in order to obtain larger sample sizes.
Metrics studied	— Machine-scored items — Human-scored item — Response times per item type
Related AQI category	Validity

Introduction

The GED Testing Service performed a large-scale, stand-alone field test (SAFT) operation in 2012. The SAFT provided an opportunity to assess the quality and performance of newly developed items from both content and statistical/psychometric perspectives. The results helped determine whether many of the assumptions about items and scoring were accurate and where improvements in item development or scoring were necessary. Data from the SAFT were used to assess item-level aspects, such as:

- Key checks
- Scoring accuracy
- Item difficulty
- Item-total correlations
- Item fit (according to the item response theory model)

Data from extended response (ER) and short answer (SA) items were also used to train and calibrate the automated scoring engine.

Method

The quality of the items was reviewed thoroughly, such that any items not performing to pre-determined standards would be excluded from future test forms. Additionally, data was aggregated to determine if certain item types were not performing well. Content specifications was also reviewed for gaps in item coverage, or content areas that were proving difficult to measure with current items or item types.

SAFT items fell into two broad scoring groups: machine-scored and human-scored. Machine–scored items included all those scored automatically by Pearson VUE's test delivery system. Human–scored items included the extended response ER and short answer SA items.

The following criteria were used to flag machine-scored items for review:

- P–value
- Point-biserial correlation
- Distribution of responses
- Percent omitted
- Percent omitted given sequence to gauge speededness

The following statistics were calculated on human-scored items:

- Item mean
- Point–biserial
- Score distribution
- Percent omitted / condition codes

Response times per item type and content area were also calculated and assessed. This information was used to help determine the amount of testing time required for the operational test. Additional information was collected from the SAFT participants after completing the study. Specifically, participants were asked about their experiences, including:

- Their level of academic preparedness for the test content
- Their computer familiarity
- The clarity of tutorial, calculator, and other test instructions
- The time allotted to complete the test modules

Findings

The SAFT yielded several different types of beneficial information. Most notable was feedback on item quality and the scoring processes. Adjustments were also made to test timing based on the response times demonstrated during the study.

2013 standardization and norming study

Study citation	GED Testing Service, *GED Test Technical Manual*, 2014.
Research study contributors	NA
Type of study	Standardization and norming study
Sample Size	1,013 current high school students
Description of sample	A two-stage sampling scheme was used. First, 100 counties were selected from a stratified probability sample. Eight strata were specified, based on the crossing of census region and rural/urban, and within strata, the probability of selection was proportional to the number of high school students. Second, all high school students within the selected counties were sought for participation, and those joining were paid to refer eligible peers. Monetary incentives and prizes were offered.

Metrics studied	— Cronbach’s alpha — Standard error of measurement — Root mean-squared error of approximation — Comparative fit index — Differential item functioning
Related AQI category	*Fairness:* Test scores can be interpreted the same way for test-takers of different subgroups. *Reliability:* An examinee’s GED test scores are stable, in that the scores are consistent, both over time and on multiple test administrations. *Validity:* GED test scores can be interpreted as a measure of examinees’ comprehension of course content, and thus, general intellectual ability.

A two-stage sampling scheme was employed to enlist current high school students into the study. In the first stage, 100 counties were selected from a stratified probability sample. Eight strata were specified based on the crossing of census region and rural/urban, and within strata, the probability of selection was proportional to the number of high school students within each county. In the second stage, all high school students within the selected counties were sought for participation, with monetary incentives and prizes being offered. Additionally, participants were paid to refer eligible peers. A total of 1,013 student participated.

Participants were randomly assigned to complete one of several core sets of test forms. These sets had one form from each content area, as well as a random subset of the remaining forms and the two GED Ready forms. In addition, each participant received a brief survey. Upon completion, students were given the opportunity to complete additional test forms for further compensation. Testing occurred under standardized conditions within authorized Pearson VUE Testing Centers.

Reliability evidence is offered in terms of internal reliability (Cronbach’s alpha) and standard errors of measurement. Cronbach’s alpha for Mathematical Reasoning (.83-.88), Reasoning Through Language Arts (.81–.84), Science (.76–.81), and Social Studies (.75–.80) were obtained for each test form. The same were obtained for subgroups based on gender and race, and results were similar across subgroups. Also, example values for conditional (on scale score) standard errors of measurement are reported. At the passing score (150), the standard errors range from 3 to 4, depending on content area, while standard errors at the honors benchmark (170) are about 4.

The main piece of validity evidence is an examination of dimensionality of the test content areas. Unidimensionality of each content area was examined by assessing the fit of single-factor analytic models to each test form. Across test forms and content areas, the root mean squared errors were below .06, and most values for the comparative fit index were greater than .95, both of which suggest that the test forms are reasonably unidimensional. As such, it is reasonable to interpret scores for each content area as being influenced by knowledge of the content areas to the exclusion of unrelated constructs.

Evidence related to fairness comes from analyses of differential item functioning, which indicate possible bias for individual items. Analyses of differential item functioning were performed based on gender, race, and ethnicity (Hispanic versus non–Hispanic) on an item-by-item basis. Items were flagged for further review based on a combination of effect size and statistical significance, and the Benjamini–Hochberg procedure was used to adjust p–values to account for multiple testing. Across test content areas, the percentages of flagged items were 5–12% based on gender, 6-25% based on race, and 1-4% based on ethnicity.

References

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

Pimentel, S. (2013). *College and career readiness standards for adult education*. Washington, DC: MPR Associates.

Pearson