datacamp
Certification

DataCamp Certification

Technical Manual

Pieter Moors, Aimée Gott and Vicky Kennedy

DataCamp
Correspondence: pieter.moors@datacamp.com

Abstract. The DataCamp Certification technical manual contains an overview of the design and development of DataCamp's career certifications. It outlines the certification process and provides evidence with respect to its validity, reliability, and fairness. This technical manual is a living document that aims to provide up-to-date information on the content and quality of DataCamp Certification.

v2023.1 Last updated on May 17, 2023

Table of contents

1 Introduction to DataCamp Certification

DataCamp Certification offers industry-recognized, career certifications that measure an individual's proficiency across the core competencies expected for each data role. That is, the achievement of a DataCamp certification means one has met or exceeded the minimum proficiency level in each of the assessed competencies. These certifications currently validate Data Analyst and Data Scientist roles at the associate and professional level, and the Data Engineer role (coming May 2023) at the associate level. To remain aligned with the needs of the job market, all certifications are developed together with a panel of data experts across a variety of industries.

DataCamp Certification is designed to be accessible any time of day, from anywhere in the world. Certification exams leverage the technology behind DataCamp assessments, enabling efficient computer-adaptive testing, and DataCamp Workspace, providing a hosted coding environment allowing candidates to complete the certification fully in the browser without the need for any local setup.

DataCamp's mission is to democratize data skills for everyone, and DataCamp Certification achieves this goal by offering industry recognized role-based certifications in an accessible and affordable way, both at the associate and professional level. The certifications are not directly dependent on DataCamp's learning content, but they are meant to formalize learning by validating the knowledge, skills, and abilities that DataCamp learners have achieved through their self-study. In this way, DataCamp Certification goes beyond statements of accomplishment awarded upon the completion of a course or track on the platform. Moreover, certifications, unlike statements of accomplishment, will expire. A certification must be renewed after a certain period of time to keep it active. This is because a certification validates that an individual has met the stated knowledge and skill level, and this knowledge and skill level must be maintained in order for a candidate to continue claiming the active certification status.

Certification programs typically offer different levels of certification, and DataCamp Certification offers an associate and professional level. Associate certifications are designed to reflect entry-level expectations in the industry, whereas the professional certification aims to reflect a skill level more commonly seen in those with at least two years of experience. Thus, obtaining DataCamp's associate certification is meant to

be a signal for employers to know the candidate has the data skills necessary for an entry role or internship as a data analyst or data scientist. In contrast, obtaining professional certification signals a proficiency level that would align with already having some experience in a data role.

2 Certification development cycle

Each DataCamp Certification is designed following the same process. Our goal is to ensure that the certification reflects the knowledge, skills and abilities (KSAs) needed to get a job in the role and level certified.

2.1 Initial job task analysis

The development of a new certification starts by conducting a job task analysis. The goal of this analysis is to understand more about the role planned to be certified, its typical responsibilities and the skills needed to perform the role at the level we plan to certify.

During this phase a range of resources is reviewed:

- **Live job posts**. Sites including indeed.com (US and UK) and LinkedIn are used to gain an overview of the requirements at varying levels. For the associate level, only jobs that include "intern", "graduate" or "junior" are considered. For the professional level, only jobs listed as mid-level e.g. "Data Scientist" are considered. Posts that suggest a combined role e.g. Data Engineer/Data Scientist are excluded, as are more senior level job posts. Job posts are considered with caution as they may not match the requirements exactly to those needed in practice. For example, a recruiter may ask for a higher level of skill than is needed.
- **Existing role frameworks**. Although not extensive, some frameworks do already exist such as the SFIA framework. SFIA is updated through an extensive consultation process, as detailed, so can be considered a reliable resource.
- **Blogs, books and learning programs**. While these resources can provide interesting input, they are considered with caution as the reasons they were put together or the inputs used to define the skills included are typically unknown.
- **DataCamp experts**. Interviews with our own data experts give valuable input at this stage.

The output of this phase is a competency matrix split by domains and levels, including tools and technologies where appropriate.

2.2 Hiring manager validation

After the domains and competencies for the role have been drafted, interviews are conducted with subject matter experts to validate them. For every role to be certified, 10-15 interviews with hiring managers are conducted. These experts are required to

have at least two years of experience in the data role with at least one year of experience hiring for the role.

Prior to conducting the interview, the following is provided:

- The draft domains and competencies
- General descriptions of how DataCamp defines levels (associate and professional)
- General descriptions of the data role (created during the initial job task analysis)

The goal of these interviews is to align the competencies with the individual's own experience in hiring for the role. During the interviews the expected tools and technologies are identified and approaches used by hiring managers during interviews to test for the competencies are discussed. Together, with the Certification Advisory Panel (see below), hiring manager interviews constitute a major source of validity evidence for the certification content developed.

2.3 Certification specification

A full specification for each certification is created from the competency matrix and the validation interviews. The specification is intended for certification developers, certification users, and certification takers. For certification developers, the specification provides a detailed description of the content breakdown, the exam structure, the item types that should be developed, and the quantities. For certification users and takers, the specification provides a detailed description of the content that will be tested and the structure of the exams.

All certification specifications can be found here:

- Data Analyst Associate
- Data Analyst Professional
- Data Scientist Associate
- Data Scientist Professional
- Data Engineer Associate

The following sections provide an overview of the information included in each specification.

2.3.1 Purpose and audience

For each certification a statement of purpose and the validity period is provided.

For example, for Data Scientist Professional:

> "This certification verifies that individuals have acquired the knowledge and skills required of mid-level data scientists, two years after entry-level.
>
> The awarded certification will be valid for two years."

Table 1: Exam structure for exam DS101, taken from the Data Science Professional Certification Specification

Domain	Technology	Number of competencies	Total items in exam
Statistical Experimentation	Python or R	7-8	15
Statistical Experimentation	Theory	15	15
Exploratory Analysis	Python or R	3-5	15

A summary of the intended users of the certification is also provided as well as their expected uses of the certification. For example:

- **Data Scientists** aiming to demonstrate that they have acquired the skills required for a mid level position through a course of training/education.
- **Hiring managers** wanting to confirm the skill level of job applicants.
- **Students** wanting to measure their achievement in a learning program.
- **Employers** wanting to confirm the current skill level of their teams.

2.3.2 Domains, competencies and KSAs

The specification provides a detailed description of what will be tested, broken down into domains, competencies, and KSAs. Where applicable, the competencies and KSAs will state the tool or technology to be used. This information is used by test developers when authoring test items.

2.3.3 Exam structure & content

The competencies are used to define the structure of both the timed and practical exams. The former consist of a series of DataCamp assessments whereas the latter is an open-ended project. The specification lists the number of timed exams and the time available for each exam. In addition, the specification provides a breakdown of the items for each domain within each exam (see Table 1).

The specification also includes the domains and competencies that will be tested in the practical exam, along with the rubric criteria that will be used in grading.

2.4 Certification advisory panel

A Certification Advisory Panel ensures that DataCamp certifications continue to reflect industry standards. The panel is comprised of individuals who have extensive experience in data related fields in both academia and industry. The panel meets quarterly to provide advice on expectations, current trends, and approaches to assessment. The composition of the panel can be found here: Introducing our new

DataCamp Certification Advisory Panel. As indicated earlier, the advisory panel constitutes a major source of validity evidence for all certification content included in the timed and practical exams.

3 Certification structure

All DataCamp certifications have the same structure. They consist of one or more timed exams and one practical exam (Figure 1).

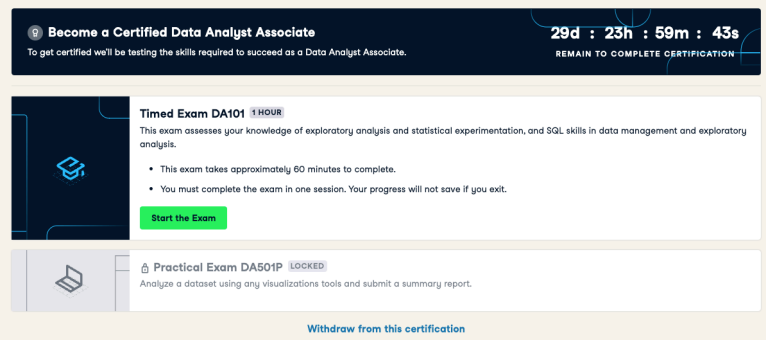

Figure 1: Example of the content a candidate needs to pass after registering for the Data Analyst Associate certification.

3.1 Timed exams

Timed exams leverage the technology behind DataCamp assessments and consist of a sequence of computer adaptive tests (CAT), each of which draws items from a pool that measures a single domain. Together, this sequence of CATs forms a single exam. Performance on each of the individual test pools is averaged to obtain a final exam score (see Section 5.2). Depending on the domains covered in an exam, candidates may be presented with any of the following three item types throughout the exam.

3.1.1 Multiple choice items

Multiple choice items include a prompt and typically four possible answers. The candidate is asked to select one of the possible answers (Figure 2).

3.1.2 Typing items

Typing items are code-based challenges where candidates need to manually type the correct solution. The submission is then evaluated by comparing the output of the

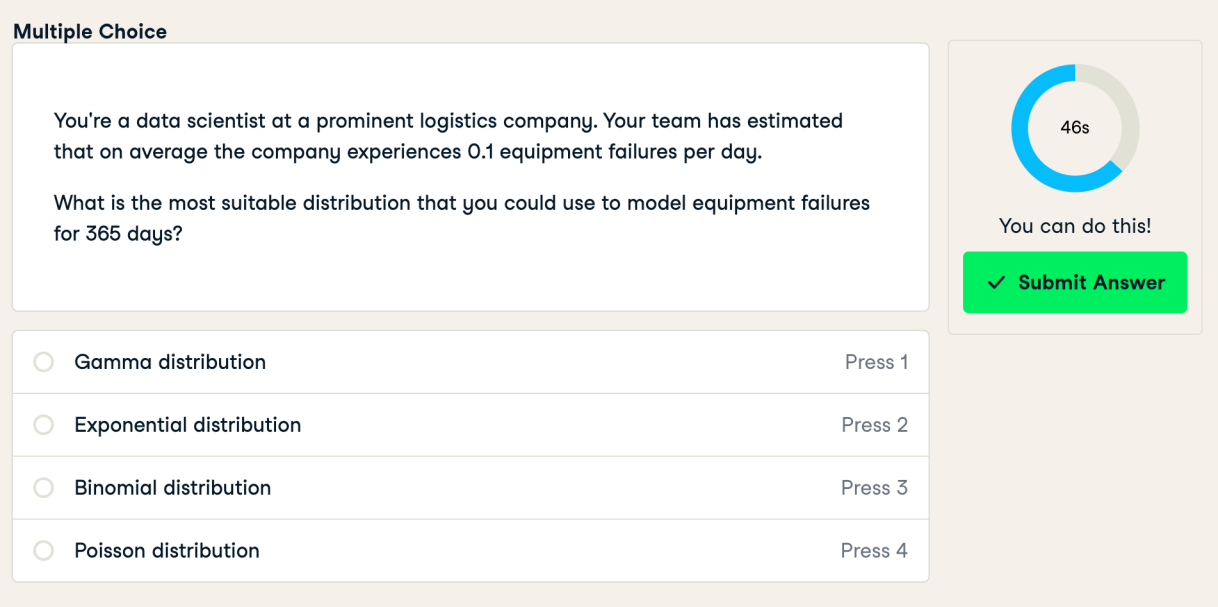

Figure 2: Example of a multiple choice item from the Learn assessment 'Statistics Fundamentals with Python'.

code submitted by the candidate with the output that the reference solution provides. In this way, different submitted solutions can yield the same output and are therefore graded as the correct solution.

Typing items contain several components. First, a prompt provides context and additional details and clarifies the task to the candidate. Second, a code block primes the task for anything required for the candidate's code to complete the task. This includes loading libraries or packages, reading in data files, or creating objects the candidate will then manipulate. In Figure 3, this would be the lines of SQL code already filled out for the candidate. In the code block, (an) empty field(s) is/are present for the candidate to complete their work per the item's instructions. For the item in Figure 3, this would be completing the INNER JOIN statement by specifying a table to be joined and the keys on which to join. Finally, and optionally, expected output is provided of the candidate's work. In general, typing items are the most difficult items in an exam as a candidate must combine technical skills, contextual information, and code syntax to successfully complete the item. Because of the complexity of the skills measured, typing items are the most frequently authored as they require more complex interactions with the candidate than other item types.

3.1.3 Fill in the blanks items

Fill in the blanks items are also code based but the candidate is asked to select the right line of code rather than typing it from scratch (Figure 4). The item structure is similar to typing items except the work the candidate produces is a selection of an option as opposed to filling out a piece of code. These items are typically used when

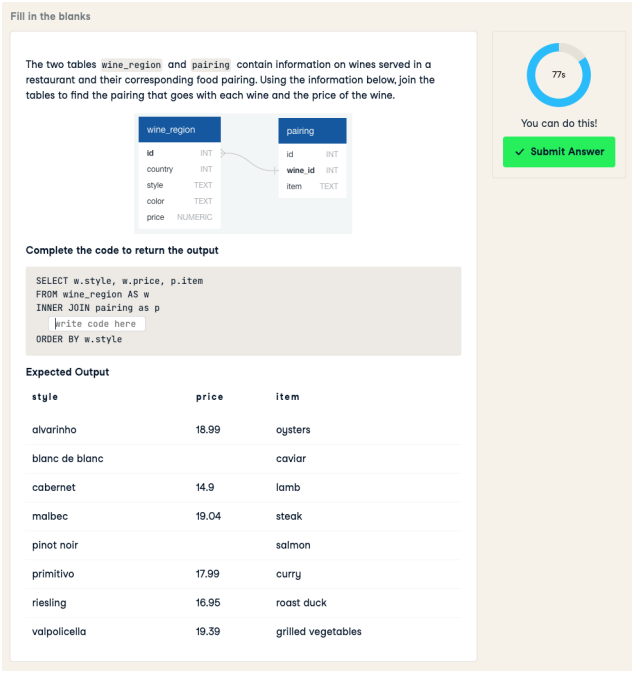

Figure 3: Example of a typing item from the Learn assessment 'Data analysis in SQL (PostgreSQL)'.

a typing item does not guarantee the same answer every time it is submitted. Fill in the blanks items include a number of distractors on top of the total number of code blocks the candidate has to submit. In some way, a fill in the blanks item is a mix of a multiple choice item and a typing item.

3.2 Practical exam

The practical exam is an open-ended project that mimics a real-world scenario. Candidates receive instructions for the practical once they pass all required timed exams.

The instructions include: - The requirements for successfully passing the project. - A section detailing the background of the business problem, including the problem statement and the questions asked. - A table containing descriptions of all the fields included in the data. - A set of tasks the candidate needs to complete in order to fulfill all requirements in the rubric. - If applicable, a sample solution setting expectations of the breadth and depth candidates need to go into to solve the business problem.

Practical exams may require the candidate to submit a written report or complete a coding project. All candidates working toward professional level certification must also submit a presentation summarizing their work to a non-technical audience. For some certifications, candidates may use any (BI) tool to analyze the data, while other

Figure 4: Example of a blanks item from the Learn assessment 'Data Management with SQL'.

Table 2: Overview of practical exam requirements for all certifications.

Certification	Analysis tool	Presentation
Data Analyst Associate	Any tool	No
Data Analyst Professional	Python/R	Yes
Data Scientist Associate	Any tool	No
Data Scientist Professional	Python/R	Yes
Data Engineer Associate	SQL	No

certifications require candidates to code in Python or R using DataCamp Workspace (Table 2).

A sample of the Data Analyst Associate practical exam can be found here. A sample solution of this exam can be found in this DataCamp Workspace.

4 Certification content development

4.1 Item development

All certification exam items go through an item development process that includes SME reviews for technical correctness as well as internal reviews to ensure that guidelines are followed, especially those related to fairness and bias. The item writing process is summarized in Figure 5. DataCamp's item writing guidelines are used across all item types and are based on the guidelines proposed by (Rodriguez, 2016).

Before they can begin writing test items, all item writers go through a rigorous hiring process and complete mandatory training. During the hiring process we are particularly looking for expertise in the field in which they will be authoring test content (minimum two years experience). Item writers are trained on the writing process and DataCamp's guidelines for writing good assessment items, and are introduced to fairness and bias in assessment and our guidelines related to this. This is mandatory training that must be completed before any item writing can begin (see Table 3 for details).

4.1.1 Timed exams

For timed exams, item writers contribute items that are written to test the KSAs defined for a domain and competency. Item writers are instructed to write items that a candidate at the level of the certification would be expected to answer. Item writers only contribute items for roles and tools/technologies for which they have prior experience. They contribute items across a range of KSAs to ensure minimum pool sizes, as specified by the certification specification, are met and that multiple item writers contribute items to each KSA.

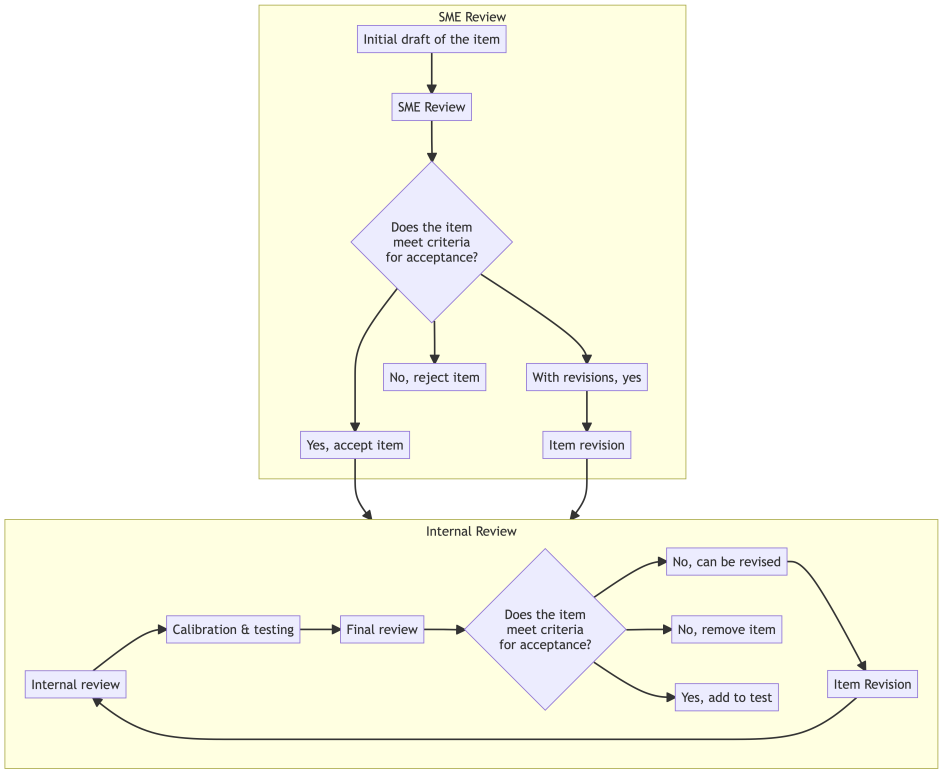

Figure 5: A flowchart of the item writing process.

Table 3: Overview of SME hiring and training procedures.

	Hiring Manager Validators	Item Writers	Graders
Experience Required	Minimum 4 years in the role to be certified and with one year in a hiring manager position for the role to be certified	Minimum 2 years in the role to be certified	Minimum 2 years in the role to be certified
Hiring/Selection Process	Confirmation of minimum experience requirements	1. Screen for experience 2. Take home test - successful authoring of test items 3. Hiring manager interview	1. Screen for experience 2. Take home test - successful grading of submissions 3. Hiring manager interview
Training	Provision of descriptors of certification levels	1. Live training on assessment and writing good practice 2. Self-paced reading and assignment using authoring tool	1. Self-paced review of training materials, rubrics and project-specific information
Ongoing Monitoring	N/A	Assessment quality metrics	Inter-grader reliability

Reviews happen at three points during the development cycle:

- **SME review**: Items are allocated to another item writer as a reviewer. This review is focused on technical correctness of the item and adherence to the KSA to be tested. Items that are deemed to be too far from the requirements may be rejected at this point.
- **Internal review**: Following revisions the item is reviewed again by a member of the DataCamp Certification Content team. This review is primarily focused on adherence to item writing guidelines including guidelines related to fairness and topics to avoid. Items may again be rejected at this point or revised.
- **Final review**: Following item calibration (where item parameters are estimated, see Section 5.1), items will be reviewed again. This review is focused on the metrics obtained during calibration including item discrimination. Qualitative feedback provided by test takers will also be reviewed. Items may at this point be removed from the test pool or further updated and re-calibrated.

4.1.2 Practical exams

4.1.2.1 Project development

Projects for practical exams are also developed with item writers, but with different requirements to the timed exams. In this case there is a more consistent structure to exams that are live at the same time to ensure consistency of testing. Item writers

additionally simulate the data sets that are used in the practical exams. It is here where most of the focus is during practical exam development.

Item writers must follow specific guidelines to ensure that candidates are able to meet the criteria defined in the rubric. This includes the structure of the data generated for the exam. Item writers must ensure that cleaning tasks required in the data are within the scope of the certification and do not prevent candidates from completing other elements of the exam, should they be unsuccessful in the data validation.

As with timed items reviews happen at three points during development, although with slightly different emphasis:

- **SME review**: Another item writer is asked to review the practical exam, with a particular focus that the data meets all requirements specified. They are additionally asked to confirm that the scenario provided is realistic.
- **Internal review**: Is again conducted by a member of the Certification Content team. As with timed items, the focus is on adherence to all guidelines.
- **Final review**: Although no calibration happens for practical exams, we aim to test all exams before they are made live in a certification.

4.1.2.2 Rubric development

Practical exams are designed with two purposes in mind:

1. To enable testing of skills that are more challenging to test in an automated manner, such as communication and business acumen.
2. To enable testing of skills required for the role in an applied manner.

The rubric for each practical exam is therefore developed with these in mind. The rubric is developed alongside a template structure for the practical exam and data requirements. Specific items in the exam are authored by item writers.

After identifying the KSAs that we intend to test in the practical exams, based on the purposes above, we then identify measurable criteria that we expect candidates to demonstrate. Where the same KSAs are tested across certifications, we will use the same criteria to demonstrate that KSA.

5 Certification delivery and scoring

5.1 Delivery

5.1.1 Timed exam

The timed exam consists of a set of one or more assessments that have been developed to measure a specific domain. For example, the DA101 exam consists of 4 assessments aimed to measure theoretical knowledge of exploratory analysis and statistical experimentation, as well as how to do data management and exploratory analysis in SQL.

Each test is delivered as a computerized adaptive test (CAT) and scored through a 2 parameter logistic item response theory model (2PL-IRT) (Van der Linden & Glas, 2010; Wainer, 2000). That is, in a given candidate's test, each item is presented conditional on the candidate's estimated ability based on the history of responses given to items presented earlier in the test. CAT procedures have been shown to generate shorter tests (Thissen & Mislevy, 2000), and more precise scores, allowing to measure many different domains in a time-efficient manner (Weiss & Kingsbury, 1984).

At the start of each test, an item of medium difficulty is presented to the candidate. Based on their response, an estimate is made of their ability and a new item is selected that is most informative to determine whether an accurate estimate of ability has been reached. To ensure the most informative items are not overexposed, the CAT algorithm selects a group of most informative items and a random item from that group is used as the next item. This method is known as randomesque selection (Kingsbury & Zara, 1989). When the domain being tested contains more than one competency, the number of items per competency is balanced by "blacklisting" items belonging to a specific competency when their maximum number has been reached (i.e., they will then no longer be considered during the test). The item selection procedure goes on until a stopping criterion is satisfied, which in our case is after the candidate has completed 15 items.

At the end of each assessment, the final ability estimate is taken as the score on that test. At the end of the exam, all scores are averaged to obtain a final exam score. This score is compared to the cut score, and a pass/fail decision is made based on whether the exam score is equal to or greater than the cut score (see Section 5.2 on how the cut score is determined). The number of assessments included in a single exam is currently always 3 or 4, resulting in a total of 45 or 60 items in each exam. On average, candidates complete a single timed exam in 45 to 60 minutes.

5.1.2 Practical exam

After passing the timed exam(s), the practical exam becomes available and candidates can start it when they wish to do so. The practical exam is self-paced but must be submitted before the deadline of the certification attempt passes (30 days after registration). Candidates receive instructions together with a personal DataCamp Workspace in which they should submit their solution. The technology needed to complete the practical exam is provided within the platform.

5.2 Scoring

5.2.1 Timed exam

All items are automatically graded as either correct or incorrect. For multiple choice and blanks items, the correct answer is known in advance and candidate input is compared with the solution to determine correctness. For typing items, the submitted code is evaluated in real-time and the output of the submitted code is compared

to the output of the reference solution. If both match, the response is graded as correct, and otherwise as incorrect. This implies that for typing items different code submissions can be graded as correct submissions.

Scoring happens based on a 2-parameter logistic item response theory model (2PL-IRT) of which the parameters have been estimated based on all valid previous user submissions, using the R package mirt (Chalmers, 2012). Maximum likelihood estimation is then used to estimate the final score based on the set of correct/incorrect candidate submissions. The ability estimate is on the latent IRT scale (centered at 0 with a standard deviation of 1), but converted to a more intuitive 0-200 scale (centered at 100 with a standard deviation of 30, where the IRT scale is capped at -3 and +3). As such, each test in the timed exam yields a score between 0 and 200. The final exam score is then averaged across those tests, and also in the 0-200 range.

Pass/fail decisions in the timed exams are based on the cut score that has been defined for each. Traditionally, standard setting methods such as the bookmark or Angoff method are used to determine a cut score. Here, we deviated from such methods as they critically depend on the notion of a minimally qualified candidate. However, all content in the timed exams has been developed with a minimally qualified candidate in mind, and applying classic standard setting approaches turned out to not be feasible. That is, independent reviewers couldn't easily identify a tipping point where a minimally qualified learner could no longer answer the questions correctly. Instead, the set of estimated item parameters within an assessment is used to determine a cut score. That is, each item pool implies an expected test score function where, for varying ability, the expected number of items responded to correctly can be determined. The expected proportion of correct responses of 0.8 was taken as performance that would be expected from any qualified candidate. This expected performance level was mapped back to the ability level that implies this expected performance. Taking this ability level as the cut score would be problematic, however, as any test has non-zero measurement error. Thus, in order to determine the cut score, a simulation-based approach was applied where the ability level that coincides with an expected performance of 80% correct was taken, and 1000s of tests were simulated. The 5% quantile of the distribution of observed scores was then used to set the cut score. This implies that a candidate whose true ability is exactly at the cut score will have a 95% chance of generating an observed score that is equal to or greater than the cut score. The average of the cut scores for all individual tests is then the cut score for the timed exam as a whole, and the average candidate score across all tests in a timed exam is compared to that cut score to determine pass/fail. By design, this procedure maximizes chances of qualified candidates to pass the timed exams, whilst incurring a (small) cost of letting non-qualified candidates pass as well. Given that the whole certification process contains many challenges, this approach is believed to contribute to a better candidate experience.

5.2.2 Practical exam

All practical exam submissions are graded by at least one human grader (see Section 9.2 on reliability and Section 10 on quality assurance), randomly drawn from a pool of human graders, against a predefined rubric. Depending on the certification,

the rubric contains 5 up to 16 different categories that graders have to mark as sufficient or insufficient. Pass/fail decisions are based on whether all components of the rubric have been graded as sufficient. A single insufficient mark results in a failing grade.

6 Fairness and acccessibility

As a platform, DataCamp is committed to serving the needs of its users by making its platform accessible for everyone https://www.datacamp.com/accessibility. DataCamp Certification is also developed with fairness and accessibility in mind. We take a universal design approach to all our certifications, seeking to maximize accessibility for all our certifications. The 24/7 availability of DataCamp Certification (no need to go to a physical testing center) and low price point (a certification can be obtained with a single monthly subscription) improve accessibility over expensive certifications that might require travel to a testing center. In addition, through DataCamp Donates (a program where free DataCamp licenses are donated to qualified organizations to people looking for work, members of disadvantaged communities, students, and nonprofit research scientists), certification is available at no cost.

During the item development process, item writers are instructed to avoid opinions and trick questions, and keep linguistic complexity at the level of the test takers. Topics that might be offensive, confusing, or require general knowledge for our candidates are actively avoided. In addition, to accommodate for users with color vision deficiency, items that require the interpretation of a graph are always designed such that they never need references to any color shown on the graph. Furthermore, font sizes on graphs should always be readable on a small screen. DataCamp supports the use of screen readers. For users with disabilities, accommodations such as extra time on the timed exams can be requested through a support ticket.

7 Certification administration

7.1 Registration

DataCamp subscribers can register for one certification at a time. After registering, candidates have 30 days to complete their certification attempt. For each part of the certification, two attempts are granted. Failing twice on any part means the candidate fails their attempt and has to wait 14 days until they can register for that certification again. This lockout period is intended to encourage candidates to review material and prepare further before attempting again. Candidates can also withdraw from their certification attempt, which will also incur a 14-day lockout period.

Table 4: Reported experience for certified users.

Experience	Proportion of certified users
Student / Less than 1 year	0.54
1-3 years	0.29
Over 5 years	0.09
3-5 years	0.08

7.2 Administration of timed and practical exam

Candidates first need to complete one or more timed exams before the practical exam becomes available. Timed exams need to be completed in order such that the second one only becomes available when the first exam is passed. When candidates click "Start exam", they are informed that the exam will take approximately 45/60 minutes to complete (depending on its length), and that they will have a maximum of 2 hours to finish the exam. As such, small breaks are allowed, but a single timed exam should be completed within the same two-hour period.

After passing the timed exam(s), candidates are free to start the practical exam when they wish. Practical exams are untimed and can be submitted anytime before the certification attempt expires. Upon submission of the practical exam, the 30-day timer is paused until the submission is graded. If a candidate fails their first attempt, the timer starts again allowing candidates to resubmit their practical based on grader feedback, assuming they still have time left.

8 Demographics of certified users

When registering for a certification, candidates are asked to fill out a few questions on their experience, background, the primary reason for pursuing certification, and whether and when they are considering a career move. The following section reports on the distribution of responses on these items for certified users.

Most certified users have less than 3 years of experience, a degree (60/40 split between non-data-related and data-related), an interest to launch a new career in a data role, and a desire to do so within 3 months.

9 Certification performance statistics

For each certification, a set of performance statistics indicative of the difficulty and quality of our timed and practical exams is monitored. In this section, pass rates for the timed and practical exams are first reported. Next, the reliability of the exam scores that drive the pass/fail decision on both types of exam is reported.

Table 5: Reported background for certified users.

Background	Proportion of certified users
I've completed a non-data-related degree	0.35
I've completed a data-related degree	0.24
I'm a student in a data-related program	0.23
I'm a student in a non-data-related program	0.09
None of the above	0.08

Table 6: Reported primary reason for certified users.

Primary reason	Proportion of certified users
I'm interested in launching a new career in a data role.	0.72
I'm seeking a promotion in my current position.	0.13
I like the challenge, but am not looking for a new position.	0.09
Other	0.05
It was required or suggested by my manager.	0.01

Table 7: Reported career move for certified users.

Career move	Proportion of certified users
Immediately	0.44
Within 3 months	0.25
Within the next year	0.19
No plans	0.11

Table 8: Pass rates for timed exams.

Certification	Exam	First attempt	First registration	All time
Data Analyst Associate	Exam DA101	0.55	0.67	0.69
Data Analyst Professional	Exam DA101	0.62	0.73	0.74
Data Analyst Professional	Exam DA201	0.82	0.92	0.93
Data Scientist Associate	Exam DS101	0.59	0.70	0.71
Data Scientist Associate	Exam DS102	0.84	0.94	0.95
Data Scientist Professional	Exam DS101	0.65	0.76	0.78
Data Scientist Professional	Exam DS201	0.69	0.84	0.87
Data Scientist Professional	Exam DS202	0.74	0.86	0.86

9.1 Pass rates

9.1.1 Timed exam

Candidates are allowed to attempt each timed exam twice within a single registration, and are allowed to re-register after a 14-day waiting period. This implies a simple pass rate of a timed exam will be a mixture of candidates who attempt the exam for the first time, the second time in the same registration, or the x-th time in subsequent registrations. Therefore, three different pass rates are reported: first attempt pass rate, pass rate in first registration, and all-time pass rate.

From the table, it is clear that the first exams in each certification are always associated with lower pass rates. This highly likely indicates that a mixture of qualified and non-qualified users attempt these. The qualified users self-select into the second exam, yielding higher pass rates. The exception seems to be Data Scientist Professional which arguably is the most difficult certification we currently offer.

9.1.2 Practical exam

Candidates are allowed to resubmit their practical exam once if they failed on their first submission. All practical exams first need to pass a technical check, and a fail on one of these (no graphics in report, no video, no audio, no screenshare, problem with the technical report, ...) does not count towards the two attempts for passing the practical exam. As for the timed exams, practical exam pass rates are reported for the first attempt, the first registration, and all-time pass rates. Here, pass rates are very much aligned for each certification, implying that the candidates who reach this stage are all at a similar ability level for the certification they are trying to obtain.

Table 9: Pass rates for practical exams.

Certification	First attempt	First registration	All time
Data Analyst Associate	0.73	0.85	0.87
Data Analyst Professional	0.75	0.84	0.88
Data Scientist Associate	0.77	0.89	0.92
Data Scientist Professional	0.71	0.84	0.87

9.2 Reliability of the timed and practical exams

9.2.1 Timed exams

For any test, reliability is an important quality metric, which is commonly assessed by quantifying test-retest reliability. The reasoning being that a test should consistently yield similar scores when a user repeats the test, even when that test consists of entirely different items — which is highly likely in a CAT setting. In our case, test-retest reliability is an ill-suited way of examining reliability since candidates who need to repeat a timed exam are a highly biased sample (i.e., the candidates that failed the first time). This yields truncated score distributions at the target score of the exam, yielding poor estimates of the true correlation between both measurement occasions. To examine reliability, the more traditional approach of quantifying a "signal-to-noise" ratio is used (i.e., the proportion of true score variance relative to the observed score variance). A marginal reliability value for each test pool is calculated, based on the following formula (Raju et al., 2007; Thissen & Wainer, 2001) - calculated through the `empirical_rxx` function of the `mirt` package (Chalmers, 2012):

$$\widehat{\rho}_{xx'} = \frac{\text{VAR}(\widehat{\theta})}{\text{VAR}(\widehat{\theta}) + \text{SE}(\widehat{\theta})^2}$$

where $\text{VAR}(\widehat{\theta})$ is equal to the variance of the estimated true scores and $\text{SE}(\widehat{\theta})^2$ is the variance of average observed standard error across candidates (SEM). The table below shows both reliability and standard error of measurement. The former ranges between 0 and 1 whereas the latter is on the 0-200 scale we report scores on on the DataCamp platform. All reliabilities are > .7 and most are > .8 indicating good reliability of the reported scores. One could argue the standard errors are a bit on the high side (e.g., the largest SEM value is 23.8 implying that the expected variability around a reported score is +- 24 points for this pool[1]). The final score on each exam is an average of 3 or 4 pools however, implying that the SEM of the total scores will range somewhere between 9 and 12 (i.e., $\text{SEM}_{\text{total score}} = \frac{\text{SEM}_{\text{exams}}}{\sqrt{\text{number of pools}}}$), which would

[1] Note that we are reporting average reliability and SEM values here, whilst IRT allows us to report conditional reliability and SEM values across the ability scale. This implies that for some parts of our scale, reliability and SEM will be better, and for other parts the values will be worse. We have opted for a single metric here for that sake of being able to efficiently summarize the quality of our timed exams.

Table 10: Empirical marginal reliability and standard error of measurement for all timed exams.

Pool name	Timed exam	Reliability	SEM
Statistical Experimentation: Theory	DA101, DS101	0.79	18.03
Exploratory Analysis: Theory	DA101	0.75	20.16
Programming For Data Science: R	DS102, DS201, DS202	0.80	18.90
Programming For Data Science: Python	DS102, DS201, DS202	0.80	19.60
Exploratory Analysis: Python	DA201, DS101	0.77	19.80
Exploratory Analysis: R	DA201, DS101	0.90	19.15
Data Management: Python	DA201, DS102, DS201, DS202	0.88	21.38
Data Management: R	DA201, DS102, DS201, DS202	0.80	20.37
Statistical Experimentation: Python	DA201, DS101	0.83	17.28
Statistical Experimentation: R	DA201, DS101	0.82	19.38
Model Development: R	DS102, DS201	0.81	19.37
Model Development: Python	DS102, DS201	0.69	24.49
Exploratory Analysis: SQL	DA101	0.86	15.81
Data Management: SQL	DA101, DS201, DS202	0.84	16.23

translate to a SEM value of .3 - .4 on the original IRT scale. Since these values are often used as stopping criteria for variable-length tests, our timed exams are sufficiently reliable and achieve satisfactory SEM (Babcock & Weiss, 2009).

9.2.2 Practical exams

Practical exams are graded by human graders against a rubric specific to each certification. Each grader assigns a sufficient/insufficient mark to each of the rubric categories and when all categories are marked as sufficient, the candidate passes the practical exam. In order to monitor grader reliability, a small percentage of practical exams is randomly assigned to two graders, such that overlapping grading data for each pair of graders is available. Many inter-grader reliability metrics like Cohen's kappa are designed to calculate agreement between two graders or whenever more than two are allowed, the data needs to be complete (i.e., each graders needs to have graded each exam). In our case, only a subset of practical exams is graded by two graders, out of a pool of more than two graders. Hence, the data is sparse rather than complete. In addition, many metrics can not properly deal with high agreement due to chance (i.e., pass rates for our practical exams are on the high side, so it is expected that graders agree by chance). Thus because more than two graders grade practical exams and those graders don't all grade the same practical exams, the generalized version of Gwet's gamma coefficient to track inter-grader reliability is used Gwet (2014). This coefficient works well in situations of high agreement, and the generalized version can handle any number of graders as well as missing data. For this, we rely on the agreement R package (Girard, 2018). The coefficient typically ranges between 0 and 1, but can be negative in case graders grade in opposite ways. For the classic kappa metric, Landis & Koch (1977) denote values > .6 and <= .8 as "substantial" agreement and anything above .8 as "almost perfect" agreement. For

Table 11: Inter-grader reliability for all rubric categories, across certifications.

Rubric category	DA Assoc	DA Pro	DS Assoc	DS Pro
Data Validation (1)	0.70	0.78	0.81	0.79
Data Visualization (1)	0.96	0.90	0.80	0.88
Business Metrics (2)	0.80	NA	NA	0.83
Business Metrics (1)	0.86	NA	NA	0.92
Communication (2)	0.91	NA	NA	0.91
Business Focus (2)	0.93	NA	NA	0.91
Business Focus (1)	0.93	NA	NA	0.93
Communication (1)	0.96	0.93	0.93	0.95
Data Visualization (3)	0.99	0.95	0.97	0.97
Model Evaluation (2)	0.99	0.96	NA	NA
Model Fitting (1)	0.96	0.99	NA	NA
Model Evaluation (1)	0.99	0.97	NA	NA
Data Visualization (2)	0.99	0.99	0.97	1.00
Model Fitting (2)	1.00	0.98	NA	NA
Model Fitting (3)	1.00	0.98	NA	NA
Business Focus (3)	0.99	NA	NA	0.98

coefficient alpha, Krippendorff (2004) recommends achieving at least .8 and accepting .667 as the minimally acceptable value. For the practical exams, inter-grader reliabiliy is aspired to be higher than .8 but any value above .6 is considered as acceptable, be it a signal for improvement. The table below summarizes inter-grader reliability for all individual rubric categories, per certification. Apart from a few categories, nearly all values exceed .8 and frequently .9, indicating good overall reliability with respect to the individual rubric categories.

9.3 Content validity

In line with the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), validation of certification depends mainly on content-related evidence (see Chapter 11 on Workplace Testing and Credentialing). In line with Standard 11.3 and 11.13, certification development starts with a job task analysis and validation through hiring manager interviews resulting in the certification specification. In addition, the Certification Advisory Panel is regularly consulted to provide advice on expectations, current trends, and approaches to assessment to continuously monitor and improve all certifications that are offered.

10 Quality assurance

Given its continuous availability and the flexibility for candidates to complete the exams at any time within the registration period, actively monitoring key metrics is highly important to ensure the certifications remain at the highest level of quality. In order to do this, a monitoring system has been set up to flag anomalies.

10.1 Timed exams

For the timed exams, three sets of metrics are relied upon. Monitoring item bank security is important to identify whether items are over- or underexposed or whether the timed exams are too similar (i.e., have too much overlap), increasing the likelihood that parts of our item bank are compromised.

The quality of items is mainly tracked through the item parameter values and user feedback. Modeling items with an IRT model has the benefit that their parameters can be used as a quality metric as well, whereas user feedback is critically important to flag issues that have not been captured by item writers or the certification content team during the item development process (e.g., typos in items).

10.1.1 Item bank security

Item exposure and test overlap metrics are actively monitored to ensure none of the items in the timed exams have a high chance of being overexposed. Item exposure rate is defined as the proportion of tests an item has appeared in. In line with Way (1998), average item exposure rates are aimed to be kept at or below 15%. The table below shows that average item exposure for nearly all pools is at or below the target value. In comparison, test overlap metrics are also tracked to quantify whether candidates are seeing many of the same items between each other (test overlap rate) or when they repeat the timed exam (repeat overlap rate). Both metrics are very similar, and a bit higher than the recommended value of 20%. It remains an active focus to increase item bank sizes such that both exposure and overlap rates will significantly decrease over the next months.

10.1.2 Item parameters

A 2PL IRT model is used and hence each item is described by its difficulty and discrimination. Item discrimination is used to monitor whether there are any poorly performing items. As a cut-off value, a discrimination value of .7 is used and an item is automatically flagged and submitted for review when it drops below that value. At the pool level, the proportion of low-discrimination items is aimed to stay below 5%.

Table 12: Average item exposure and test overlap rates.

Pool name	Timed exam	Item exposure	Test overlap	Repeat overlap
Statistical Experimentation: Theory	DA101, DS101	0.16	0.33	0.34
Exploratory Analysis: Theory	DA101	0.14	0.31	0.32
Programming For Data Science: R	DS102, DS201, DS202	0.09	0.32	0.34
Programming For Data Science: Python	DS102, DS201, DS202	0.10	0.31	0.31
Exploratory Analysis: Python	DA201, DS101	0.15	0.32	0.32
Exploratory Analysis: R	DA201, DS101	0.12	0.28	0.31
Data Management: Python	DA201, DS102, DS201, DS202	0.12	0.25	0.28
Data Management: R	DA201, DS102, DS201, DS202	0.09	0.27	0.32
Statistical Experimentation: Python	DA201, DS101	0.11	0.28	0.28
Statistical Experimentation: R	DA201, DS101	0.08	0.18	0.23
Model Development: R	DS102, DS201	0.10	0.26	0.27
Model Development: Python	DS102, DS201	0.13	0.32	0.31
Exploratory Analysis: SQL	DA101	0.13	0.30	0.31
Data Management: SQL	DA101, DS201, DS202	0.13	0.25	0.25

10.1.3 Item feedback

During timed exams, candidates can provide feedback on any item they are completing. They can submit technical problems, problems with the questions or answer choices given, or spelling and grammar issues with the item. The technical issues are sent to the engineering team to improve the quality and uptime of the timed exams, while the substantive feedback is important to the certification content team to continuously improve items or remove items from the item bank.

10.2 Practical exams

There are only a few live versions of each practical exam project in place at any one time. For such an open-ended assignment, metrics such as item exposure and overlap rates do not make sense. Hence, the focus is mainly on comparability of pass rates across projects within a certification. Furthermore, double-graded submissions allow to monitor both inter-grader reliability and evaluate the quality of the double-graded submissions.

10.2.1 Practical projects

The quality of the practical exam projects is primarily evaluated through the submission and pass rates. That is, if, for the same certification, a project has a higher/lower submission rate compared to another, this is indicative of a difficulty mismatch. Next, if submission rates are equal, pass rates are examined as well. They are expected to be roughly equal as well. If this is not the case, this would be indicative of a mismatch between the submitted projects and the graders who need to evaluate them. Resolving mismatches in pass rates is usually done by aligning the graders on the rubric for the project under consideration. In addition, candidates and/or graders sometimes flag issues themselves (similar to item feedback in timed exams), which are taken on by the content team.

10.2.2 Grading QA flow

A QA flow, triggered on each submission, monitors the quality of the grades submitted by the human graders. All submissions are graded by a "primary" grader. A "secondary" grader is assigned with a probability of 20%. These overlapping grades are used as input to quantify inter-grader reliability. Last, a "secondary/tertiary" QA grader is assigned with a probability of 20%. This latter assignment is independent of the former and is primarily used to evaluate whether all graders are grading against the rubric. Any submission can thus have 1-3 graders. Upon submission of the grades, absolute agreement is determined. If any of the graders disagree on at least one of the rubric categories, the submission is sent to another grader who needs to resolve the disagreement and submit the final grade.

11 Concluding remarks

DataCamp Certification aims to provide industry-recognized, role-based certifications that measure proficiency across the core competencies necessary for each (data) role. In this technical manual, evidence for the validity, reliability, and fairness of DataCamp's certifications is provided. The technical manual documents how each certification is designed, how its content is developed, and how they are administered and scored. This technical manual is a living document and will be updated each time either a new certification is added or when any of the components are updated to further improve the validity, reliability, or fairness of the certifications. This technical manual was last updated at 05/17/2023.

References

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. AERA.

Babcock, B., & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable-length CATs are not biased. Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing, 14.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48, 1–29.

Girard, J. M. (2018). Agreement (version 0.0.0.9003). https://github.com/jmgirard/agreement.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. The British Journal of Mathematical and Statistical Psychology, 61(1), 29–48.

Gwet, K. L. (2014). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters (4th ed.). Advanced Analytics.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education, 2, 359–375.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. Human Communication Research, 30(3), 411–433.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 159–174.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. Applied Psychological Measurement, 31(3), 169–180.

Rodriguez, M. C. (2016). Selected-response item development. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), Handbook of test development. Routledge.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), Computerized adaptive testing: A primer. Routledge.

Thissen, D., & Wainer, H. (Eds.). (2001). Test scoring. Lawrence Erlbaum Associates Publishers.

Van der Linden, W. J., & Glas, C. A. (Eds.). (2010). Elements of adaptive testing (Vol. 10, pp. 978–970). Springer.

Wainer, H. (2000). Computerized adaptive testing: A primer. Routledge.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. Educational Measurement: Issues and Practice, 17(4), 17–27.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21, 361–375.