Codility_

AI for Business Technical Report Version 2.0



Last Update: October 2026

CONFIDENTIAL AND PROPRIETARY

Table of Contents

Executive Summary 5
1. Introduction 7
  1.1 Background 7
  1.2 Current approaches to measuring AI-related skills 7
  1.3 Purpose, audience, and scope 8
  1.4 Evidentiary framework 8
2. Construct Definition and Theoretical Foundations 10
  2.1 Definition of the construct 10
  2.2 The work-sample rationale 10
  2.3 Automated scoring of constructed responses 11
  2.4 Anticipated threats to validity 11
3. The Skills Model 12
  3.1 Structure of the model 12
  3.2 Development of the model 12
  3.3 Applicability to business roles 13
4. Assessment Design and the Task Library 14
  4.1 Components of a task 14
  4.2 The assistant as part of the instrument 14
  4.3 Task formats 14
  4.4 Tiers 15
  4.5 Composition of the standard library 16
5. Content-Related Evidence 17
  5.1 Method 17
  5.2 Distribution of criterion-to-skill links 17
  5.3 Representation of the AI category 18
  5.4 Domain expertise 18
  5.5 Checks on the accuracy of the linkage 18
6. Development Process and Scoring Verification 19
  6.1 The development pipeline 19
  6.2 Release criteria 20
  6.3 Study 2: Library-wide scoring verification 20
  6.4 Study 4: Expert review of the short-form tasks 21
  6.5 Consequences for development standards 22
7. Scoring Method and Agreement with Expert Judgment 24
  7.1 The scoring engine 24
  7.2 Study 1: Concordance with expert human judgment 24
  7.3 Follow-up: synthetic concordance and correction of criteria 26
  7.4 Partial credit 26
  7.5 Comparison with a holistic language-model judgment 27
  7.6 Robustness to construct-irrelevant response strategies 27
8. Field Study in Live Hiring (Study 3) 28
  8.1 Setting 28
  8.2 Method 28
  8.3 Results 29
  8.4 Observations reported by the partner's raters 30
  8.5 Discussion and limitations 30
9. Reliability and Generalizability 31
  9.1 Sources of measurement error 31
  9.2 Correlation between tasks 31
  9.3 Projected consistency of multi-task composites 31
  9.4 Operational score distributions 31
10. Subgroup Differences and Adverse Impact (Study 5) 33
  10.1 Analytic framework 33
  10.2 Method 33
  10.3 Summary of results 34
  10.4 Results by gender 35

© 2026 Codility. All rights reserved.

10.5 Results by race and ethnicity....35
10.6 Discussion and limitations....36
11. Candidate Reactions....37
11.1 The survey instrument....37
11.2 Results....37
11.3 Differences between women and men....38
11.4 Candidate comments....38
12. Language, Test Security, and Assistance from External AI....40
12.1 Language....40
12.2 Test security and the data-access design....40
12.3 Assistance from external AI....41
13. Measuring the AI Interaction (Study 6)....42
13.1 Planned capabilities....42
13.2 Study 6: Transcript-evaluation methods....43
14. Intended Use, Limitations, and Research Agenda....47
14.1 Uses supported by the current evidence....47
14.2 Uses requiring further or local evidence....47
14.3 Standard setting and norms....47
14.4 Limitations of the current evidence....48
14.5 Research agenda....48
Appendix A. Study Summaries....50
Appendix B. Task Evidence Register....53
Appendix C. Methods Notes....55
C.1 Agreement statistics....55
C.2 Study 3 cross-tabulation and statistics....55
C.3 Adverse impact methods....55
C.4 Spearman–Brown projections....56
References....57
Version History....59

© 2026 Codility. All rights reserved.

Executive Summary

AI for Business is a family of constructed-response work samples in which a candidate uses an embedded, task-specific AI assistant to investigate a realistic business problem and then submits a short written work product. The work product is scored by a deterministic natural language processing (NLP) engine against explicit criteria defined for each task. This report documents the construct the assessments are intended to measure, the design of the tasks and of the scoring method, and the empirical evidence accumulated between April and October 2026. Consistent with current professional guidance, the evidence is organized as an argument for a specific interpretation and use of scores (Kane, 2013; AERA, APA, & NCME, 2014). The argument proceeds through four questions: whether scores are derived accurately from responses, whether they generalize beyond the particular task a candidate completed, whether they correspond to expert judgments of performance in an operational setting, and whether their use is consistent with the fairness obligations described in the *Uniform Guidelines on Employee Selection Procedures* (1978).

Six studies contribute to that argument (Table 1). In a scoring concordance study (Study 1), 200 synthetic responses to a single scenario were scored by the engine and by three expert raters. Engine scores agreed with the expert consensus at levels conventionally described as good to substantial (ICC = .77; quadratically weighted κ = .76; r = .80), and disagreements of more than one point on the three-point scale occurred in 1% of responses; agreement was weakest for partially correct responses. A follow-up analysis against a language-model rater found that several criteria which behaved correctly on exemplar text failed on newly written responses, and those criteria were corrected. A library-wide verification study (Study 2) applied a five-test protocol to each of the 53 tasks then available. Twenty-five tasks satisfied all five tests on the first pass, and the remainder exhibited specific, documented defects, most often susceptibility to answers produced by an external AI model from the task description alone. That susceptibility was concentrated in two task formats, and the finding has since shaped development standards and a hardening program. A field study in live hiring (Study 3), conducted with a large U.S. telecommunications company, compared engine scores for 96 applicants with ratings assigned independently by the partner's trained subject-matter expert. Mean engine scores increased monotonically across the five levels of the expert's rating scale, from 37% to 87%, and the rank-order association was substantial (Spearman ρ = .72). Absolute agreement could not be evaluated as an equivalence test, because only one expert rated the responses and the engine scores were grouped into uncalibrated bands for comparison.

Three further studies address the newer short-form tasks, subgroup differences, and the measurement of the AI interaction itself. An expert review of the twenty short role-agnostic tasks (Study 4), conducted under decision rules fixed before data collection, found construct fit rated 4 or 5 on a five-point scale in 20 of 22 ratings and led to targeted scoring changes in eight tasks. A subsequent synthetic validation of the same tasks found a median false-accept rate of approximately 25%, which is the principal reason these tasks are not yet recommended as stand-alone hurdles in high-stakes decisions. An analysis of 1,006 operational sessions (Study 5), 330 of which included self-reported demographic data, found negligible differences between women and men ($g = -0.04$ to 0.15) and small standardized differences among racial and ethnic groups ($|g| \leq 0.37$). At selective hypothetical cut scores, selection rates for White and for Hispanic or Latino candidates fell below four-fifths of the rate for Asian candidates in several analyses. These results are reported in full, with the sampling limitations that constrain their interpretation. A program of interim studies of transcript evaluation (Study 6), conducted on transcripts of AI-assisted software engineering rather than on AI for Business sessions, found that language-model judgments of transcripts were highly repeatable across runs (ICC = .85 to .95) but unrelated to an AI-generated measure of session success. In the first comparison with an expert's ratings, a holistic judgment agreed moderately with the expert (ICC =

© 2026 Codility. All rights reserved.

.63), whereas methods built from behavioral indicators agreed poorly (ICC = .11 to .26). No transcript-derived score is used in decisions.

Supplementary analyses document the linkage between scoring criteria and the Codility Engineering Skills Model (626 criteria carrying 1,862 criterion-to-skill links, 69% of criteria tagged to at least one AI skill). They also report a correlation of r = .41 between two tasks of different formats completed by the same candidates (n = 716), and candidate reactions collected through a redesigned post-test survey. Taken together, the evidence supports interpreting scores as indicators of AI-mediated problem solving on the tasks sampled, and using them to rank-order candidates within multi-task batteries, with human review of mid-range scores. The evidence does not yet establish criterion-related validity, equivalence between engine and human scoring on authentic responses under a two-rater design, or normative interpretations. Section 14 describes the uses the current evidence supports, the uses it does not, and the studies planned to address the remaining questions.

Table 1. Studies contributing to the validity argument

Study	Period	Design	Sample	Principal result
1. Scoring concordance	Apr 2026	Engine scores compared with the consensus of three expert raters; follow-up against a language-model rater	200 synthetic responses, one scenario	ICC(2,1) = .77, r = .80; disagreement beyond one scale point in 1% of responses
2. Library-wide scoring verification	Jun 2026	Five-test protocol executed in the live candidate environment	53 tasks; 8 reviewers; 264 completed runs	25 of 53 tasks passed all five tests on first pass; exposure to outside AI concentrated in two formats
3. Field study in live hiring	Jun–Sep 2026	Custom task built from a job analysis; engine compared with one trained partner expert	96 applicants	Mean engine score rose from 37% to 87% across expert rating levels; ρ = .72
4. Expert review of short-form tasks	Jul–Sep 2026	Synthetic concordance; expert review under fixed decision rules; in-platform attempts	20 tasks; 3 reviewers	20 of 22 construct-fit ratings ≥ 4 of 5; median synthetic false-accept rate ≈ 25%
5. Subgroup differences and candidate reactions	Jul–Oct 2026	Uniform Guidelines selection-rate analyses; standardized mean differences; survey	1,006 sessions; 330 with demographics	Gender differences negligible; ethnic differences |g| ≤ .37; impact ratios < .80 at selective cuts
6. Transcript-evaluation methods (interim)	Jul–Oct 2026	Language-model judgments of engineering transcripts in repeated runs; comparison with one expert's ratings	200 and 300 external sessions; 92 and 233 candidates; 61 expert-rated	Run-to-run ICC .85 to .95; agreement with the expert ICC .63 (holistic) vs. .11 to .26 (indicator-based)

© 2026 Codility. All rights reserved.

1. Introduction

Generative AI tools have moved rapidly from specialist use into ordinary business work, and employers increasingly ask whether candidates can use them to reach accurate, defensible conclusions rather than merely fluent ones. Credentials, interviews, and self-report measures do not observe this capability directly, and knowledge-based measures capture what candidates know about AI rather than what they do with it. This section describes the measurement problem that motivated the AI for Business assessments, places the assessments among current approaches to measuring AI-related skills, and sets out the purpose, scope, and evidentiary framework of the report. The central premise is that the capability of interest is best measured by observing it in a standardized simulation of the work, and that such a measure must then be supported by the same kinds of evidence required of any selection procedure.

1.1 Background

Teams in sales, marketing, legal, and operations now use AI tools to build their own automations and to complete tasks that once required specialist support (Anthropic, 2026). The practical consequence for organizations is that the quality of an employee's collaboration with AI has become a component of job performance in roles well beyond software engineering. A tool that summarizes a dataset or drafts a recommendation adds value only when the user asks well-formed questions, recognizes when the output is wrong or incomplete, and takes responsibility for the conclusion that is finally communicated. These are the behaviors that distinguish productive use from uncritical reliance, and they are not directly observed by the methods most commonly used to select candidates for non-technical roles.

The AI for Business assessments were developed to observe these behaviors under standardized conditions. Each task presents a realistic business problem, an evidence base that can be explored only through an embedded AI assistant, and a request for a short written work product. Because the work product is scored, and because the correct conclusion depends on information the candidate must elicit and verify, the score is intended to reflect the quality of the result the candidate achieves with the assistant.

1.2 Current approaches to measuring AI-related skills

In October 2026 the Assessment Science team reviewed approximately 30 vendors offering assessments described as measuring AI skills, AI fluency, or AI readiness. The review classified offerings by what they observe and by the evidence their publishers make available (Table 2). Most offerings rely on self-report or disposition measures, or on judgment items about AI that candidates answer without using AI. Developer-oriented platforms have extended into business roles, but where they score AI use automatically, they do so mainly within coding tasks. A small number of newer vendors place a live AI inside a business work sample. None of the offerings reviewed was accompanied by published reliability, criterion-related, or adverse-impact results for an assessment in which a live AI participates in the task. Published guidance on evaluating AI-based employment tools sets out the questions that vendors of such tools should be prepared to answer (Campion & Campion, 2026), and those questions inform the organization of this report.

Table 2. Approaches to measuring AI-related skills (October 2026 review)

Approach	What is observed	Evidence published by vendors

© 2026 Codility. All rights reserved.

Self-report and disposition profiles	Attitudes, self-rated proficiency, learning agility	General questionnaire studies
Judgment items about AI	Choices about prompting, verification, and responsible use, made without using AI	None identified
Developer platforms	AI use in coding tasks; some role-titled tests with embedded AI, scored on outputs or by an interviewer	Bias audits claimed; no results identified
Live AI in business work samples (other vendors)	AI use in a simulated workspace	None identified
AI for Business	AI-mediated problem solving that produces a scored business deliverable	Studies 1–6 (this report)

Conversations held between May and October 2026 with talent-acquisition, people-analytics, and industrial-organizational psychology staff in customer organizations, and with other assessment providers, were consistent with the review. The request heard most often was to observe a candidate using AI on a realistic task. I-O and legal reviewers placed particular weight on the transparency of the scoring method and asked specifically for criterion-related and adverse-impact evidence.

1.3 Purpose, audience, and scope

The report is intended for the professionals whom customers ask to evaluate assessment instruments, including industrial-organizational psychologists, talent assessment leaders, and legal and compliance reviewers. It is organized to support review against the AERA/APA/NCME *Standards for Educational and Psychological Testing* (2014) and the *Uniform Guidelines* (1978). The report covers the AI for Business category specifically. Platform-level material, including the development of the Codility Engineering Skills Model, integrity and proctoring safeguards, platform scoring conventions, and the evidence base for Codility's established assessment types, is documented in the *Codility Technical Manual* (Codility, 2026a) and is referred to rather than repeated.

1.4 Evidentiary framework

Validity is a property of the interpretation and use of scores, not of a test itself, and the case for an interpretation is assembled from several kinds of evidence (AERA et al., 2014; Messick, 1995). Kane's (2013) argument-based approach makes this explicit by identifying the chain of inferences that leads from an observed performance to a decision, and by asking what evidence supports each link. The interpretation proposed here is that an AI for Business score indicates a candidate's proficiency at AI-mediated problem solving on realistic business tasks, and that scores aggregated across tasks can inform selection decisions when combined with other information. Table 3 maps each inference in that argument to the sections and studies that bear on it and indicates the current state of the evidence.

Table 3. Inferences in the validity argument and the evidence that bears on them

Inference	Claim	Evidence in this report	Current status
Scoring	Scores are derived accurately and consistently from responses, and responses that should not earn credit do not	Study 1 (concordance), Study 2 (verification), Study 4 (short-form tasks), Section 7	Supported for synthetic responses on one scenario; verification identified and documented defects

© 2026 Codility. All rights reserved.

Generalization	Scores on the tasks administered generalize to the broader domain of comparable tasks	Cross-task correlation and score distributions (Section 9)	Initial evidence from one task pair; multi-task batteries required
Extrapolation	Scores reflect the capability as it is exercised in the job	Construct definition (Section 2), content linkage (Section 5), field study (Study 3, Section 8)	Content linkage documented; convergent evidence with expert ratings in one live setting
Implication and decision	Decisions based on scores are appropriate and consistent with fairness obligations	Subgroup analyses (Study 5, Section 10), candidate reactions (Section 11), intended use (Section 14)	Early operational evidence; criterion-related evidence and operational cut-score analyses pending

© 2026 Codility. All rights reserved.

2. Construct Definition and Theoretical Foundations

Any interpretation of a score depends first on a clear statement of what the score is intended to measure and of what it is not intended to measure. This section defines the construct targeted by AI for Business, explains why a work sample was chosen to operationalize it, summarizes the research on automated scoring of constructed responses, and identifies the principal sources of construct-irrelevant variance that the design must control. The construct is AI-mediated problem solving in a business context. A work sample operationalizes the construct directly, and work samples have an established record as predictors of job performance. The principal cost of the format, the task-specific variance characteristic of performance assessments, is addressed through multi-task measurement and examined empirically in Section 9.

2.1 Definition of the construct

The construct is defined as the ability to obtain useful and accurate results from AI tools while performing a realistic job task. It comprises four related behaviors. The first is formulating questions that elicit the information a problem requires. The second is probing an evidence base through the AI instead of accepting a summary at face value. The third is validating what the AI returns against source data and against the internal logic of the problem. The fourth is communicating a conclusion that the evidence supports. The definition is consistent with research on AI literacy, which characterizes competence with AI as multidimensional, spanning knowledge, application, evaluation, and ethical judgment (Long & Magerko, 2020; Ng et al., 2021). It corresponds most closely to the AI Fluency and AI Collaboration skills of the Codility Engineering Skills Model (Section 3) and to public-sector definitions of AI application as a competency, which pair knowledge of AI capabilities and limitations with skill in using AI to produce useful and accurate work products.

Three boundaries of the construct are relevant to interpretation. First, the scored tasks are not tests of knowledge about AI technology; where such knowledge is of interest, separate multiple-choice knowledge sets are available. Second, they are not intended as measures of general writing ability: the written product is evaluated principally for the substance of its conclusion, supporting evidence, and reasoning. Grammar and spelling are not scored, although a small number of short-form criteria concern tone, length, or presentation. Third, for role-agnostic tasks, they are not intended to measure occupational knowledge beyond what the task materials supply. Each boundary implies a potential source of construct-irrelevant variance, considered in Section 2.4.

2.2 The work-sample rationale

Each task is a simulation in which the candidate receives a role, a business problem, an evidence base of data extracts, reports, and correspondence, and an embedded AI assistant, and must deliver a short recommendation, briefing, or set of findings. Work samples are established predictors of job performance. The most recent meta-analytic review places their operational validity at .33 (Sackett et al., 2022). That estimate draws on the updated meta-analysis of Roth et al. (2005) and supersedes the earlier estimate of .54 (Schmidt & Hunter, 1998). The rationale for the format rests on behavioral consistency: performance on a faithful sample of the work is expected to predict performance on the work itself. Work samples are also

© 2026 Codility. All rights reserved.

among the selection methods that applicants tend to perceive most favorably (Hausknecht et al., 2004), a consideration examined for these tasks in Section 11.

Several AI for Business scenarios resemble take-home analytical exercises that predate generative AI. A design premise of the series is that effective collaboration with an AI assistant allows a candidate to complete in twenty to thirty minutes an analysis that would otherwise require several hours, and that the time limits therefore favor candidates who use the assistant well. This premise has not yet been tested directly, for example by comparing performance and completion times with and without the assistant.

2.3 Automated scoring of constructed responses

A substantial literature supports automated scoring of constructed responses in selection. NLP-based scoring of candidate essays has been shown to approach the reliability of trained human raters (Campion et al., 2016). Machine-learning approaches to scoring constructed responses and interview content have shown acceptable psychometric properties and meaningful relationships with job-relevant outcomes (Koenig et al., 2023; Liff et al., 2024). Campion (2026) concludes that such methods can meet legal and professional standards for selection when they are transparent and properly validated. Williamson et al. (2012) describe a framework for evaluating automated scoring along several dimensions: the construct relevance of the scoring features, agreement with human scores, relationships with external criteria, generalizability across tasks and populations, and fairness. The evidence in Sections 5 through 10 is organized to address each of these dimensions.

2.4 Anticipated threats to validity

Messick (1995) distinguished two broad threats to the validity of score interpretations: construct-irrelevant variance, in which scores are influenced by characteristics other than the intended construct, and construct underrepresentation, in which the measure omits important aspects of the construct. Five threats were identified as most relevant to this assessment design, and each is examined in a later section.

The first is **reading load and English proficiency**. Tasks are delivered in English and require reading scenario descriptions and materials, so proficiency in English could influence scores independently of the construct (Section 12). The second is **occupational knowledge** beyond what the materials supply, which would advantage candidates with specific domain experience on tasks intended to be role-agnostic (Section 5). The third is **assistance from external AI tools**, which could allow a candidate to obtain credit without exercising the construct within the task environment (Sections 6 and 12). The fourth is **response style relative to the scoring exemplars**. Because several criteria compare responses with exemplar answers, a correct response phrased unusually could receive less credit than its substance warrants (Section 7). The fifth threat concerns construct underrepresentation: because only the final work product is scored, behaviors such as delegation and iteration are observed only indirectly (Sections 5 and 13).

© 2026 Codility. All rights reserved.

3. The Skills Model

The constructs that AI for Business scores are defined in the Codility Engineering Skills Model (ESM), the competency framework maintained by Codility's Assessment Science team. A skills model contributes content-oriented evidence: it establishes that the constructs being measured are defined, bounded, and representative of a domain. It does not by itself establish the validity of any assessment built from it. This section summarizes the structure of the current revision (v2.1), the evidence from which it was developed, and the extent to which a model built largely with engineering subject-matter experts can be applied to business roles. The model's artificial intelligence category, introduced as a first-class domain in v2.1, provides the primary definitions for the AI skills scored here. Its applicability to business roles rests on evidence gathered for individual tasks, not on the model's development panels alone.

3.1 Structure of the model

ESM v2.1 organizes 30 skills and 180 subskills into five categories: Technical, Artificial Intelligence, Problem-Solving, Business, and Technology Stack. Each skill is a labeled construct with a behaviorally oriented definition, and each subskill describes a more specific way in which the skill is exercised. The Artificial Intelligence category comprises the five skills summarized in Table 4. AI for Business scores behavior primarily within the Artificial Intelligence and Problem-Solving categories, together with selected Business skills such as Communication.

Table 4. The Artificial Intelligence category of the Codility Engineering Skills Model (v2.1)

Skill	Definition (abridged)	Subskills
AI Fluency	Working effectively, efficiently, ethically, and safely with AI by understanding what it is, how it works, and what it can and cannot do	AI Foundations, AI Failure Modes, Context Engineering, AI Data Literacy, Context Hygiene
AI Collaboration	Working with AI as a partner across the full interaction loop, from delegation through output refinement, including supervising autonomous agents	AI Task Delegation, AI Iteration, AI Output Validation, Agent Orchestration
AI Responsibility	Ensuring AI is used fairly, transparently, and accountably, treated as an active competency	Algorithmic Fairness, AI Explainability, AI Accountability, AI Ethics
AI Governance	Operating within the regulation, policies, and standards that govern AI use	AI Regulation, AI Policy, Human Oversight, AI Risk Management
AI Engineering	Designing, building, deploying, and maintaining AI and ML systems (primarily technical roles)	Machine Learning, LLM Development, AI Architecture, MLOps, AI Evaluation, AI Security, AI Cost Optimization

3.2 Development of the model

The model was developed through successive layers of evidence, not in a single exercise. Its foundation consolidated more than thirty published skill and competency frameworks, including the IEEE SWECOM model and the Software Engineering Body of Knowledge. These were refined and rated by panels of more than forty-five internal and external subject-matter experts, and a skill was retained only with the agreement

© 2026 Codility. All rights reserved.

of two-thirds of the experts. A work analysis of AI-assisted engineering work followed. It combined the research literature, O*NET work activities, and direct observation of practitioners using AI tools, and a work activity was linked to a skill only where an independent expert panel confirmed the linkage. A multinational survey of respondents in twenty-four countries then added quantitative evidence of criticality, asking whether each skill was critical and required on entry rather than merely important. The expansion of the model to AI readiness drew on a structured review of the AI-literacy literature (Long & Magerko, 2020; Ng et al., 2021) and on interviews and focus groups with technology leaders representing more than 120 years of collective experience.

Revision v2.1 re-anchored the model through a three-phase method consistent with published guidance on competency modeling (Campion et al., 2011). The first phase surveyed the current skills landscape. It drew on discipline-wide bodies of knowledge (SWEBOK v4), leveled skill frameworks (SFIA 9), government and labor-market taxonomies (O*NET, Lightcast), regulation and standards (the NIST AI Risk Management Framework, the EU AI Act, and ISO/IEC 42001), and recent observational evidence on how AI is used at work. In the second phase, subject-matter experts named, bounded, and organized the constructs the survey identified. In the third phase, the revised model was circulated to thirty-five experts in a confirmation survey in which any category, skill, or subskill could be flagged as missing, misnamed, misdefined, or misplaced. The model was adjusted only where feedback was corroborated across sources. The full methodology is reported in the ESM technical report (Codility, 2026b).

3.3 Applicability to business roles

During the AI-readiness expansion, experts classified each AI skill according to whether it applies to engineering roles specifically or to the general workforce. That classification identifies the AI skills that are meaningful to measure in candidates outside engineering. However, the panels that confirmed the model's constructs were drawn largely from engineering, so the model's development evidence speaks more directly to technical roles than to business roles. For AI for Business, the relevance of the model to business work must therefore be supported task by task; it cannot be assumed from the model as a whole. Three sources of task-level evidence are available: the criterion-level linkage described in Section 5, the independent expert review of the short-form tasks described in Section 6.4, and, for custom deployments, the customer's own job analysis, as in the field study described in Section 8.

© 2026 Codility. All rights reserved.

4. Assessment Design and the Task Library

This section describes the instrument itself: the components of a task, the role of the embedded assistant, the task formats and tiers, and the composition of the standard library as of October 2026. Two features of the design have particular consequences for interpretation. The AI assistant is part of the standardized stimulus rather than a tool the candidate brings, so its behavior must be specified and held constant across candidates. And the score evaluates a work product that the candidate may have produced partly with the assistant's help, so the construct is the quality of the result obtained with the AI, not the candidate's unaided writing.

4.1 Components of a task

Each task has five components. The scenario establishes the candidate's role, the business problem, the requested deliverable, and the response guidelines. The materials form the evidence base, such as data extracts, logs, financial records, policy documents, and correspondence, and are authored with relevant signals and plausible distractors deliberately embedded. The embedded assistant, Cody, is the candidate's principal interface to those materials; since August 2026, candidates can also open read-only views of the attached data files (Section 11.4). The response editor is a writing environment without coding controls, in which text formatting is not evaluated. The scoring key, which the candidate does not see, decomposes the correct answer into discrete criteria.

4.2 The assistant as part of the instrument

Each task ships with a behavior specification that governs what the assistant knows, what it reveals and under what conditions, and what it declines to do. The specification serves two purposes. It standardizes the interaction across candidates, so that differences in outcomes reflect differences in how candidates use the assistant rather than differences in what the assistant was willing to provide. It also insulates the task from changes in the underlying language model, which would otherwise alter the stimulus over time. Whether the specification fully achieves the second purpose across model versions has not yet been evaluated empirically.

Candidates may move text produced by the assistant into their response, and the full conversation is captured and attached to the candidate report. The score therefore evaluates the submitted product, whether the candidate wrote it unaided or with the assistant's help. This is consistent with the construct definition, which concerns the quality of the result a candidate obtains and verifies, but it means that the score alone does not indicate how much of the product the candidate composed. The interaction measures described in Section 13, including the proportion of the submission generated by the AI, are intended to make that information available to reviewers.

4.3 Task formats

The library uses six task formats, summarized in Table 5. Four formats are used in the medium and hard tiers and differ in the reasoning they demand: weighing competing courses of action, auditing an AI-generated report against source data, extracting specified findings from a multi-source evidence base,

© 2026 Codility. All rights reserved.

and writing a prompt that must reliably produce an accurate work product. The two short formats, introduced in 2026, reduce the reading load and time required while retaining an AI-mediated deliverable.

Table 5. AI for Business task formats

Format	Candidate activity	Primary behaviors elicited	Scored criteria
Decision briefing	Weighs competing courses of action, each supported by a plausible stakeholder position, and recommends the one the evidence supports	Evidence-based reasoning under ambiguity; resistance to anchoring on authority	9–12
AI output audit	Verifies an AI-generated report against source data, identifying embedded errors such as unsupported figures, methodological errors, and misleading framing	Validation of AI output; calibrated skepticism	9–12
Research drilldown	Extracts a defined set of findings from a multi-source evidence base under time pressure	Delegation to, and interrogation of, the assistant	5
Prompt engineering	Writes a prompt that must reliably generate an accurate, audit-ready work product	Precise specification; anticipation of failure modes	20
Lean (short form)	Writes one prompt, or corrects one flawed AI draft for tone, omission, factual error, jargon, or length	Context engineering; output validation; communication	5
Inbox (short form)	Answers a workplace email using two text sources queried through the assistant, whose first answer is designed to be misleading	Triage, verification, and judgment among competing demands	8

4.4 Tiers

The formats are organized into three tiers that differ in length and depth (Table 6). The short tiers were developed in response to customer feedback that the medium and hard tasks were too long for some high-volume and early-career contexts. One customer set a benchmark of approximately five minutes per task, beyond which, in that customer's experience, candidates were lost from the process. Other customers asked that shorter tasks continue to resemble actual work. The short tiers trade content depth for time: they score five or eight criteria, compared with five to twenty (median 12) in the medium and hard tiers, which limits the precision of any single short task and makes their use in combination with other tasks more important (Sections 9 and 14).

Table 6. Library tiers

Tier	Approximate time	Tasks	Design	Origin
Lean	5 min	10	One screen of context; a single short deliverable	Customer benchmark of about five minutes per task
Inbox	10 min	10	Two text sources; one email reply	Customer requests for short tasks that still resemble work

© 2026 Codility. All rights reserved.

Medium and hard	20–50 min	51	Multi-source evidence base; designed trap conclusions; longer reasoning chains	Original series for professional hiring

Custom tasks are developed through the same pipeline from a customer's own job-analytic data. The custom task examined in Study 3 had a 75-minute limit.

4.5 Composition of the standard library

As of October 2026 the standard library documented in this report comprises 71 work samples: 51 medium and hard tasks across nine role families and 20 short role-agnostic tasks (Appendix B). The broader catalog, including custom and specialized tasks, contains approximately 80 work samples. Table 7 summarizes the library's characteristics. During the 2026 preview period, coding-oriented interface elements were removed from the candidate experience in response to enterprise feedback, including program-output controls, technology lists, and coding terminology in instructions, and the longest task descriptions were shortened. One description, for example, was reduced from 817 to 458 words. These changes were made to remove cues irrelevant to the construct and to reduce reading load.

Table 7. Characteristics of the AI for Business library (October 2026)

Attribute	Status
Standard library	71 work samples (51 medium and hard; 20 short role-agnostic), listed in Appendix B
Full catalog	Approximately 80 work samples, including custom and specialized tasks
Role families	Customer success, customer support, data analysis, finance, marketing, product management, product marketing, sales, and general workforce (role-agnostic)
Difficulty levels	Elementary, easy, medium, hard
Time limits	Approximately 5 to 50 minutes; custom tasks up to 75 minutes
Scored criteria per task	5 to 20 (median 12 among medium and hard tasks)
AI knowledge sets	14 multiple-choice sets (137 items); four sets revised in October 2026 to model-neutral wording
Delivery	Codility Screen, Interview, and Skills Intelligence; available since Public Preview in July 2026
Language	English

© 2026 Codility. All rights reserved.

5. Content-Related Evidence

Content-related evidence concerns the relationship between what an assessment asks candidates to do and the construct it is intended to measure (AERA et al., 2014). For constructed-response tasks scored by explicit criteria, the most informative level at which to document that relationship is the individual criterion, because the criteria define what earns credit. This section describes how the criteria are linked to the Skills Model, reports the distribution of those links across skills, and examines where the linkage is strong, where it is sparse, and how its accuracy has been checked. The linkage indicates that the score is weighted primarily toward problem solving and AI-related skills. Within the AI category, it concentrates on output validation and context engineering, while delegation, iteration, and responsible-use skills are sparsely represented in the written-product scores.

5.1 Method

The linkage matrix for the medium and hard library was compiled in June 2026, after the tasks had been built, by applying a structured, partly automated tagging workflow to each task's scoring criteria. For each criterion, the workflow assigned as many as three ESM skills, each at a defined proficiency level, together with a written rationale citing the relevant ESM definition and the content of the criterion. The resulting matrix covers 626 criteria across the 56 medium and hard tasks in the June 2026 library and contains 1,862 criterion-to-skill links. The number of criteria per task ranges from 5 to 25, with a median of 12. The short-form tasks were tagged in a comparable manner. Because the tags were produced by a single workflow applied to finished scoring code, their accuracy depends on that workflow and on the review its output received. Independent verification by reviewers who had not seen the tags has so far been conducted only for the short-form tasks (Study 4), and an equivalent review of the medium and hard library is part of the research agenda (Section 14).

5.2 Distribution of criterion-to-skill links

Table 8 shows the distribution of links across ESM categories. Problem-solving skills account for 34% of links, AI skills for 30%, business skills for 27%, and data skills within the Technical category for 8%. At the level of criteria, 69% carry at least one AI-skill tag, 82% carry at least one problem-solving tag, and 99% carry one or the other. The assigned proficiency levels are concentrated at the intermediate (826 links) and advanced (649) levels, with fewer at beginner (256), expert (90), and novice (41), which is consistent with tasks designed for working professionals.

Table 8. Criterion-to-skill links by ESM category (626 criteria, 56 tasks)

ESM category	Links	Share	Most frequent skills (links)
Problem-Solving	641	34%	Critical Thinking (496), Systems Thinking (145)
Artificial Intelligence	565	30%	AI Output Validation (256), Context Engineering (167), AI Failure Modes (80)
Business	509	27%	Domain Expertise (345), Communication (155)
Technical: Working with Data	147	8%	Data Exploration (87), Working with Data (43)
Total	**1,862**	**100%**	

© 2026 Codility. All rights reserved.

5.3 Representation of the AI category

Within the AI category, links are concentrated in three subskills: AI Output Validation (256 links), Context Engineering (167), and AI Failure Modes (80). Other subskills are represented sparsely: AI Iteration (23), AI Accountability (15), AI Task Delegation (13), AI Data Literacy (4), AI Foundations (3), Human Oversight (2), and Context Hygiene (2). The pattern is consistent with what a written work product can reveal. Whether a candidate validated the assistant's output, and whether the information requested was sufficient, leave traces in the conclusion and evidence the candidate reports. How a candidate delegated, how often the candidate iterated, and whether oversight was exercised are properties of the process, and are visible mainly in the interaction transcript. The written-product score should therefore be interpreted as underrepresenting the process components of AI Collaboration (threat five in Section 2.4). Measurement of those components is the purpose of the work described in Section 13. Skills of AI Responsibility and AI Governance are largely outside the scope of the current task formats, and dedicated formats for them are under consideration.

5.4 Domain expertise

Domain Expertise is the single most frequently tagged business skill (345 links, 19% of all links). In role-based tasks this is intended: occupational knowledge is part of the work of, for example, a financial planning analyst reconciling a month-end close. In role-agnostic tasks, any reliance on occupational knowledge would be construct-irrelevant (threat two in Section 2.4). A customer review of the role-agnostic tasks identified Domain Expertise tags on tasks intended to be role-agnostic, and an audit of those tasks for reliance on knowledge not supplied by the materials is planned.

5.5 Checks on the accuracy of the linkage

Three checks on the accuracy of the linkage are available to date. The first was a design review of the inbox tier before release. It found that the draft criteria mostly carried non-AI tags (critical thinking, communication, and teamwork), and that two of the five criteria in a typical task were etiquette or empathy markers, so that roughly 40% of each score reflected email conventions rather than the AI skills each task was intended to measure. Every inbox scoring key was re-keyed before release, so that AI-skill criteria carry between 50% and 75% of each task's score and presentation is limited to a single criterion (12.5%). The second check was the independent expert review in Study 4, in which reviewers named the skills each task required before viewing the assigned tags. Where two reviewers disagreed with a task's tags, the task was rebuilt so that the tagged skill determined the outcome. In one such task, whose description had disclosed the flaw candidates were meant to detect, 22 of 40 weak responses now fail the detection criterion that they had previously passed by default. Disagreements from a single reviewer led to targeted revisions of individual criteria or wording. The third check was the review of the custom task in Study 3 by the partner's subject-matter experts, who examined the task and its model answer against their own job analysis before the task was used in hiring.

© 2026 Codility. All rights reserved.

6. Development Process and Scoring Verification

The scoring inference requires that scoring rules be applied as intended, that correct responses receive credit, and that responses which should not earn credit do not. Statistical agreement with expert raters on one scenario (Section 7) does not establish these properties for every task in a library, so two further studies examined them directly. Study 2 applied a structured five-test protocol to every medium and hard task in the live candidate environment, and Study 4 combined synthetic concordance analysis with expert review of the short-form tasks. This section first describes the development pipeline and release criteria that both studies evaluated, then reports the studies. On first pass, Study 2 found defects in 28 of the 53 tasks it examined. The verification studies also showed that exposure to answers generated by external AI models depends systematically on task format, and their findings have changed the standards under which new tasks are built.

6.1 The development pipeline

Each task is developed through a six-stage pipeline with a human review checkpoint at every stage (Table 9). The first stage, the generation brief, fixes the role, difficulty, format, correct conclusion, and the trap conclusions that an insufficiently careful analysis would reach, and is the stage at which construct intent is most directly determined. Scoring generation follows an eight-step procedure. The brief is parsed into a blueprint, a map of valid answers is constructed, and the answer is decomposed into concept units. Each unit is classified as required, alternative-valid, or a negative indicator. Lexical domains are then generated, four to six exemplar responses covering the valid reasoning paths are written, thresholds are calibrated against distributions of correct and incorrect responses, and the key is subjected to adversarial validation. Tasks that fail validation return to the scoring stage; the evidence base is held fixed and only the scoring logic is revised.

Table 9. Development pipeline stages and outputs

Stage	Primary output	Focus of human review
1. Generation brief	Role, difficulty, format, correct conclusion, designed trap conclusions	Construct intent and job relevance
2. Scenario generation	Business problem, candidate instructions, assistant behavior rules	Realism, clarity, fairness of framing
3. Materials generation	Evidence base with embedded signals and distractors	Internal consistency; solvability from the data
4. Scoring generation	Criteria, exemplar and graded-incorrect responses, calibrated thresholds	Coverage and defensibility of the answer space
5. Assistant configuration	Final behavior specification	Standardization; absence of answer leakage
6. Validation gate	Automated test results and content sign-off	Release decision

© 2026 Codility. All rights reserved.

6.2 Release criteria

The automated stage of the validation gate applies the criteria in Table 10. Human reviewers separately evaluate task content against a rubric covering job relevance, clarity, and freedom from construct-irrelevant demands.

Table 10. Automated release criteria

Check	Criterion
Exemplar (golden) responses	Score ≥ 80% (short-form tasks: 100%)
Concise but correct responses	Score ≥ 80%
Responses reaching a trap conclusion	Score ≤ 40%
Keyword-stuffed responses	Score ≤ 40% (short-form tasks: 0%)
Prompt injection, refusals, and unedited placeholders	Score 0% (validity gate)
Unedited starter response	Score 0%
Graded incorrect responses	Scores decrease monotonically with response quality
Criterion groups	No group fails entirely on a correct response

6.3 Study 2: Library-wide scoring verification

Purpose and method. Study 2 examined whether the scoring of each medium and hard task behaved as specified when the task was taken in the live candidate environment, as opposed to the development environment in which release criteria are first applied. Between June 17 and 25, 2026, eight reviewers from the Assessment Science, content, and product teams applied five tests to each of the 53 tasks then in the library. The tests submitted, respectively, the exemplar answer, the unedited starter response, the exemplar with required components removed, off-task or generic AI-generated content, and an answer produced by an external AI model from the task description alone (Table 11). Reviewers recorded each score and a pass or fail judgment against the expected behavior. The protocol defined failure on the fifth test as a passing-level score without specifying a numeric threshold, and classifications of intermediate scores (approximately 33% to 60%) were not fully consistent across reviewers. Of the 265 planned runs, 264 were completed.

Table 11. Study 2 protocol and first-pass results (53 tasks)

Test	Expected behavior	First-pass result
1. Exemplar solution	Scores at the top of the range	49 of 53 (92%)
2. Unedited starter	Scores zero	46 of 53 (87%)
3. Component removal	Score decreases in proportion to what is removed	52 of 53 (98%)
4. Off-task AI content	Earns no credit	42 of 53 (79%)
5. External AI answer	Does not reach a passing-level score without the in-task data	31 of 52 run (60%)
All five tests		**25 of 53 tasks (47%)**

© 2026 Codility. All rights reserved.

Results. The component-removal test behaved as expected in 52 of 53 tasks, indicating that scores decline in proportion to missing substance and do not hinge on isolated keywords. The exemplar test failed in four tasks. Two of these failures involved alternate exemplars that scored below target, and the review also traced exemplar failures to scenario company names that had been revised after the exemplars were written, a form of drift the test is designed to detect. In seven tasks the unedited starter received credit (between 8% and 33%), and in eleven tasks off-task content did; eight of the eleven off-task failures occurred in prompt-engineering tasks. Results of the external-AI test differed markedly by format (Table 12). Research drilldowns and AI output audits, in which the answer resides in the data, were rarely solvable from the description alone: one of 14 drilldowns and two of the 14 audits for which the test was run reached a passing-level score. Decision briefings and prompt-engineering tasks were more often solvable: 11 of 15 and 7 of 9, respectively. Reviewers also observed that an external model could sometimes suggest useful questions to pose to the in-task assistant even when it could not answer the task itself.

Table 12. Study 2 results by task format

Format	Tasks	Passed all five tests	External-AI test not passed
Research drilldown	14	12	1
AI output audit	15	10	2
Decision briefing	15	3	11
Prompt engineering	9	0	7
Total	**53**	**25**	**21 (1 not run)**

Discussion. The format effect has a plausible mechanism. A decision briefing lists its candidate courses of action in the task description, and the scenario framing often makes it possible to infer which option the evidence is likely to favor without seeing the evidence. A prompt-engineering deliverable can be partly composed from the description's statement of what the prompt must accomplish. In both formats, then, the description carries information about the answer that the data-access design was intended to withhold. The finding is therefore a threat of the third kind identified in Section 2.4, and it is concentrated in identifiable formats rather than spread across the library. The exemplar and starter findings were defects in individual task builds, not in the scoring method. Study 2 is engineering and content verification. It establishes whether specified scoring behavior holds, which is distinct from the psychometric evidence in Sections 7 through 10. The results describe the task versions of June 2026; re-test results for remediated tasks are not reported here.

6.4 Study 4: Expert review of the short-form tasks

Synthetic concordance. Before expert review, each of the ten lean tasks was evaluated by scoring 50 synthetic responses with the engine and classifying the same responses independently with a language-model judge. The judge was used only for validation and plays no part in operational scoring. The targets were concordance of at least 80% between the engine's and the judge's accept or reject decisions, a false-accept rate of at most 10%, and a false-reject rate of at most 15%. Seven of the ten tasks met all three targets at their operating thresholds. Of the remaining three, one was stricter than the judge by design, a concision task that produced no false accepts; in the second, the residual discrepancies involved responses the judge itself rated as borderline; and the third, a task requiring the correction of a wrong date and time, exceeded the false-accept target at its operating threshold (12%) and fell below the concordance target at the stricter threshold (76%). The procedure also revealed an error in the judge: its rubric initially

© 2026 Codility. All rights reserved.

treated the flawed draft's values as the target of correction, causing correct revisions to be rated as failures. The rubric was corrected before any results were used.

Expert review. Study 4 then examined all twenty live short-form tasks between August 21 and September 1, 2026. Three assessment scientists took part, and all three worked six anchor tasks chosen because their evidence was weakest or their risk highest. Two reviewers rated construct fit and the defensibility of the scoring key on five-point scales across 16 tasks, yielding 22 ratings on each scale. The third reviewer, who had developed the tasks, contributed in-platform attempts and defect notes. Four tasks did not receive a human construct rating and were cleared on measurement evidence alone. Two procedures protected the independence of the ratings. Reviewers named the skills a task required before viewing the assigned tags, and they judged each scoring key only from the description the candidate sees, not from the scoring code. The decision rules were fixed before data collection. A task would be held if the median defensibility rating was 2 or below, or if one reviewer flagged an unfair trap and another confirmed it. A task would be edited if the median construct-fit rating was 3 or below, if two reviewers disagreed with its difficulty label, or if the median time judgment exceeded twice the label. Any expert attempt scoring below 75% would trigger investigation of the task's scoring gates, and tags would be re-derived wherever two reviewers disagreed with them. The results are summarized in Table 13.

Table 13. Study 4 results

Measure	Result
Construct fit (1–5)	20 of 22 ratings at 4 or 5
Scoring-key defensibility (1–5)	19 of 22 ratings at 4 or 5
Fairness flags	3 in 32 reviews (one ambiguous request, two concerns about traps), each resolved through revisions to criteria or to candidate-facing disclosure
Defects identified in the preceding review round	53 by inspection; 7 further by a follow-up measurement pass
In-platform findings	An editor formatting issue that could cause a correct reply to score zero, corrected during the review
Release decision	All 20 tasks cleared, 8 after targeted scoring changes
False-reject rate after revisions	0% to 9.1% across tasks

Discussion and limitations. The ratings indicate that, in the judgment of experienced reviewers, the short-form tasks require the skills they are tagged to measure and that their scoring keys can be defended to a candidate who challenges a score. Study 4 is nonetheless a screening study: it reduces the risk of a subsequent validation study but does not substitute for one. It did not include a candidate sample, so it cannot provide item statistics or agreement between engine and human ratings of candidate responses. Before the review, engine agreement with an AI rater on synthetic responses had ranged from ICC = .23 to .75 across the twenty tasks. A later synthetic validation found a median false-accept rate of approximately 25%, meaning that weak responses were accepted more often than the target allows, with four of the twenty tasks reaching ICC of .70. This later estimate, which covers all twenty tasks, differs from the earlier lean-tier analysis, in which eight of ten tasks met the false-accept target and seven met all three targets, and it is the estimate on which the recommendations in Section 14 rest. Agreement indices of this kind are attenuated when response quality varies little, as it does on easy tasks, which is why accept or reject concordance was adopted as the primary calibration check for these tiers. The targets nonetheless remain a median false-accept rate below 10% and ICC of at least .70 on at least fifteen of the twenty tasks, and work toward them is scheduled to begin in October 2026. Until those targets are met and human ratings of

© 2026 Codility. All rights reserved.

candidate responses are available, the short-form tasks are recommended as components of broader batteries rather than as stand-alone hurdles (Section 14).

6.5 Consequences for development standards

The findings of Studies 2 and 4, together with the design review of the inbox tier, led to several changes in the standards under which tasks are now built. A validity gate assigns a score of zero to prompt injection, refusals, non-answers, and unedited placeholders. The starter text must be a bare instruction that scores 0%, so that no scored content is supplied to the candidate. Content checks were made sensitive to negation and to the scope of a claim, so that quoting a flawed draft, or arguing against a required correction, does not earn credit for the correction. Candidate-facing descriptions of how a reply is assessed were aligned with what the criteria actually score, so that no criterion depends on an expectation the candidate has not been told about. Finally, a hardening program was defined for the formats most exposed to external AI, as described in Section 12.3.

© 2026 Codility. All rights reserved.

7. Scoring Method and Agreement with Expert Judgment

This section describes how the scoring engine evaluates a response and reports the principal evidence that its scores correspond to expert judgment. Evaluation against calibrated human raters is the established standard for automated scoring (Williamson et al., 2012). The question it addresses is whether the engine assigns scores that a trained expert would assign to the same responses. In Study 1, the engine agreed with the consensus of three expert raters at levels conventionally described as good to substantial, with large disagreements rare and accuracy lowest for partially correct responses. The section also reports an internal comparison with a holistic language-model judgment, which bears on the choice of a deterministic method. The evidence currently rests on synthetic responses to a single scenario, and the extensions required to generalize it are identified.

7.1 The scoring engine

The written work product is scored by an NLP engine that executes a set of task-specific test cases, analogous to the test-case architecture used across Codility's platform. Criteria use three families of methods. Lexical criteria detect required conclusions, figures, and concepts, and include checks for negation and for the scope of a claim. Semantic-similarity criteria, based on sentence embeddings (SBERT), compare the meaning of the response with the task's set of exemplar responses. In many medium and hard tasks, overlap criteria based on TF-IDF vocabulary similarity and ROUGE-1 key-point coverage compare the response with the same exemplar set. The short-form tasks rely on lexical criteria applied behind a validity gate and a core-decision gate. Criterion results are aggregated into a task-level percentage, and the reviewer receives a plain-language account of which criteria were met.

The engine contains no generative model. Scores are a deterministic function of the response text, and every criterion can be inspected, reproduced, and challenged. These properties are relevant to the auditability expected of selection procedures and to emerging regulatory frameworks such as the EU AI Act. The similarity and overlap criteria carry a specific risk, identified as threat four in Section 2.4: a correct response phrased very differently from the exemplars could receive less credit than its substance warrants. The development pipeline mitigates this risk in three ways. It requires four to six exemplars covering distinct valid reasoning paths, it requires that concise but correct responses score at least 80%, and, for the short-form tasks, it monitors false-reject rates. The magnitude of the risk for responses from candidates who are not native speakers of English has not been estimated.

7.2 Study 1: Concordance with expert human judgment

Method. In April 2026, the engine was evaluated against expert judgment on a scenario in which candidates diagnose an 18% decline in an executive dashboard metric that followed a pricing and packaging change, and recommend one of four courses of action to the chief executive (Codility, 2026c). The correct response recognizes that the decline is largely a measurement artifact, produced by a change in the metric's definition and the filtering of automated traffic, rather than a deterioration in the health of the business. Two hundred synthetic responses were written to span the range of quality: clearly correct responses with complete reasoning, partially correct responses with gaps in diagnosis or in the specificity of the recommended action, and clearly incorrect responses representing three characteristic misdiagnoses. The engine scored each response with ten criteria, five lexical and five comparing the response with reference

© 2026 Codility. All rights reserved.

text (semantic-similarity and vocabulary-overlap criteria). Three expert raters scored each response on three dimensions, recommendation accuracy, diagnostic reasoning, and action specificity, each on a three-point scale. Each rater's dimension scores were averaged and rounded to a single rating, and the three ratings were averaged and rounded to form the consensus criterion. The engine's percentage score was mapped to the same scale (0% to 33% = 1, 34% to 66% = 2, 67% to 100% = 3). Agreement was evaluated with ICC(2,1) under a two-way random-effects, absolute-agreement model, quadratically weighted κ, ordinal Krippendorff's α, and Pearson's r. Classification accuracy was evaluated by treating the consensus as the reference category and computing precision, recall, and F1 for each category.

Results. The raters agreed closely with one another (Krippendorff's α = .91; pairwise ICC(2,1) and weighted κ from .92 to .95), providing a stable criterion. Agreement between the engine and the rater consensus is summarized in Table 14. By the benchmarks of Koo and Li (2016) and Landis and Koch (1977), agreement was good to substantial, and Krippendorff's α exceeded the .667 minimum suggested for tentative conclusions (Krippendorff, 2004). Agreement between the engine and individual raters was lower than agreement among the raters (ICC(2,1) = .70 to .79, compared with .92 to .95), and the engine's correlations with individual raters ranged from .75 to .82. Exact category agreement with the consensus was 69%. Disagreements of one category occurred in 30% of responses and disagreements of two categories in 1%. The classification results show where the disagreements lay. All 46 responses in the consensus low category were placed in the low category by the engine (recall = 1.00), and of the 78 responses the engine placed in the high category, 71 were rated high by the consensus (precision = .91). The engine was less accurate for the remaining high responses, 37 of the 110 being placed in the middle category (recall = .65), and for the middle category itself, in which it placed 21 of 44 responses correctly, 16 lower, and 7 higher (F1 = .41). Macro-averaged F1 across the three categories was .67. The engine's mean rating was 2.07, compared with 2.29 to 2.39 for the raters. The study tentatively attributed this conservative tendency to strict thresholds on the semantic-similarity criteria, which credited only language close to the exemplars.

Table 14. Study 1: Engine–human scoring concordance (N = 200 responses)

Comparison and metric	Value	Interpretation
Human–human agreement (Krippendorff's α)	.91	Above the .80 threshold; stable criterion
Human–human pairwise ICC / weighted κ	.92–.95	Excellent
Engine vs. human consensus, ICC(2,1)	.77	Good (Koo & Li, 2016)
Engine vs. human consensus, weighted κ	.76	Substantial (Landis & Koch, 1977)
Engine vs. human consensus, Krippendorff's α	.73	Above the .667 minimum (Krippendorff, 2004)
Engine vs. human consensus, Pearson r	.80	Strong linear association
Engine vs. individual raters, ICC(2,1)	.70–.79	Moderate to good
Exact category agreement	69%	About twice the rate expected by chance on a 3-point scale
Adjacent disagreement (one category)	30%	
Disagreement beyond one category	1%	
F1 by category (low / middle / high)	.84 / .41 / .76	Weakest for partially correct responses
Macro-averaged F1	.67	
Mean rating, engine vs. raters	2.07 vs. 2.29–2.39	Engine about 0.25 points lower

© 2026 Codility. All rights reserved.

Discussion and limitations. The results support the scoring inference for responses of the kind studied: the engine separates strong from weak responses largely as the expert raters did and rarely makes large errors, but it classifies partially correct responses poorly (F1 = .41). Four limitations constrain generalization. First, the responses were synthetic rather than produced by applicants, and synthetic responses may differ from authentic ones in their phrasing and in the ways they go wrong. Second, the study covers one scenario, which is not among the tasks in the current standard library. Third, thresholds were calibrated during development, and the extent to which the reported agreement reflects calibration to these particular responses has not been estimated. Fourth, the three raters were members of the team that developed the assessment and were familiar with its intended solution. Their close agreement with one another is therefore partly a product of shared training and expectations, and agreement with raters independent of development may be lower. A replication on authentic responses not used in calibration, rated by experts independent of development, is therefore part of the research agenda. Study 3 provides a first test with authentic applicant responses on a different task (Section 8). Consistent with the study's mid-range finding, supplementary human review of mid-range scores is recommended (Section 14).

7.3 Follow-up: synthetic concordance and correction of criteria

The mid-range weakness and the conservative tendency identified in Study 1 were examined further in a follow-up analysis, which also tested a procedure for applying the same method to every task without repeated manual rating (Codility, 2026d). Fifty newly generated synthetic responses to the same scenario, ten in each of five quality strata, were scored by the engine and by a language-model rater whose ratings had been checked against the human ratings used as an anchor (α = .974). The language-model rater was used only to evaluate the engine and plays no part in operational scoring. Initial agreement was markedly lower than in Study 1 (ICC(2,1) = .49; exact agreement 50%).

Diagnosis traced the gap to defects that calibration against exemplar text had not revealed; for the overlap criteria, the calibration files were verbatim copies of the reference answers. Two semantic-similarity thresholds were so strict that newly written responses almost never passed them. A lexical-overlap criterion based on Jaccard similarity had a maximum attainable value for novel text of about .22, below its threshold of .35, so it failed every response, including the strongest. A TF-IDF criterion discriminated in the wrong direction, passing none of the ten strongest responses but three of the twenty responses in the next two strata. Two keyword entries credited observations that also appear in incorrect responses as though they signaled the correct recommendation. Seven successive changes corrected these defects: the semantic thresholds were recalibrated, the Jaccard criterion was replaced by ROUGE-1 recall, the TF-IDF criterion was removed, the keyword lists were corrected and expanded, partly with phrasings drawn from the same fifty responses, and two criteria were added, one crediting engagement with the scenario's quantitative evidence and one failing responses that explicitly recommended one of the incorrect options. After the changes, agreement with the language-model rater reached ICC(2,1) = .87, weighted κ = .87, α = .88, and exact agreement of 80%, with no disagreements of two categories. F1 was .77, .58, and .93 for the low, middle, and high categories (macro-averaged F1 = .76). The middle category remained the weakest, and the analysis concluded that further adjustment of binary criteria would not improve it.

Two features of this analysis limit the weight it can bear. The criterion was a language-model rater rather than human experts, and the same fifty responses were used both to diagnose and correct the criteria and to evaluate the result, so the final agreement is an optimistic, in-sample estimate. The analysis is reported for its methodological contribution: it showed that criteria which pass every check against exemplar text

© 2026 Codility. All rights reserved.

can fail on newly written responses, a finding directly relevant to the verification studies reported in Section 6.

7.4 Partial credit

Study 1 and its follow-up both located the engine's principal weakness in the middle of the scale and attributed it to criteria that award all-or-nothing credit. Study 1 recommended converting lexical criteria from binary to graded credit (0, 0.5, or 1; Codility, 2026c), and lexical criteria were subsequently revised to award graded credit, so that scores decline in proportion as required content is removed. This behavior is consistent with the component-removal results of Study 2 (52 of 53 tasks), although that test does not distinguish graded from binary credit. When the graded criteria were evaluated, classification improved over the binary baseline (macro-averaged F1 = .795 vs. .763), with the largest gains in the middle of the score range, where binary criteria are least informative. The binary baseline corresponds to the final result of the follow-up analysis in Section 7.3, in which a language-model rater served as the criterion. The graded criteria have not yet been evaluated against human ratings or on authentic responses; that evaluation is part of the replication of Study 1 described in Section 14.5.

7.5 Comparison with a holistic language-model judgment

The decision to exclude generative models from scoring was also examined empirically. In internal testing during 2026, a structured holistic judgment by a large language model was compared with expert ratings. The model's judgments were highly repeatable across runs (ICC = .88) but agreed with a subject-matter expert at only .09. The comparison was internal and limited in scope, and its details are not reported here, so it should be regarded as preliminary. It nonetheless illustrates a distinction that recurs throughout this report: repeatability is necessary for a defensible score but is not evidence of accuracy, since a scorer can be consistently wrong. The research on transcript evaluation reported in Section 13.2 shows the same pattern, with agreement between language-model judgments and an expert ranging from ICC = .11 to .63 depending on how the judgment was structured. Deterministic criteria, by contrast, allow accuracy to be examined one criterion at a time, as in Studies 1 and 2.

7.6 Robustness to construct-irrelevant response strategies

Scoring keys include adversarial test cases directed at known strategies for obtaining unearned credit. Responses that list required terms without connecting reasoning are capped well below a passing score. Attempts at prompt injection against the scoring pipeline are tested explicitly and receive no credit. Off-task content, including plausible generic text and AI-generated summaries of the materials, is expected to score at or near zero, a property tested in Study 2 (79% of tasks met it on first pass). Resistance to answers produced by external AI models is reported by format in Section 6.3 and addressed by the program described in Section 12.3.

© 2026 Codility. All rights reserved.

8. Field Study in Live Hiring (Study 3)

Studies 1 and 2 examined scoring under controlled conditions. The question that follows is whether engine scores correspond to expert judgment when the responses are produced by real applicants, on a task built for a real role, in an operational hiring process. Study 3 addressed that question in partnership with the selection and assessment function of a large U.S. telecommunications company. A custom task was developed from the partner's job analysis and administered to 96 applicants as a step in hiring. Each response was scored by the engine and rated independently by a trained subject-matter expert employed by the partner. Mean engine scores increased at every level of the expert's rating scale, and the rank-order association between the two was substantial. Because only one expert rated the responses, and because the engine scores were grouped into bands not calibrated to the expert's scale, the study provides convergent evidence for the scoring inference on authentic responses but not a test of equivalence between engine and human scoring.

8.1 Setting

The role was a senior analyst position in a customer-relations team that resolves escalated customer complaints, in which decisions carry reputational and regulatory consequences. The team hires periodically from applicant pools of approximately 150. Before the partnership, AI skills for this role were assessed manually: candidates completed a task with a general-purpose AI tool during a proctored video call, and a single moderator observed, collected outputs, and scored each one. The partner regarded that process as time-intensive, difficult to scale, and difficult to standardize. Because it intended to use the new assessment as an independent step in hiring, the partner set a correspondingly high evidentiary standard.

8.2 Method

Job analysis and task development. The partner's selection and assessment team conducted a formal job analysis, using structured interviews and surveys with subject-matter experts, to identify the critical skills of the role, including how AI collaboration fits into daily work. Codility developed a custom task from the results, drawing on sample documents, realistic scenarios, and job-shadowing input supplied by the partner. In the task, candidates have 75 minutes to query the assistant across six workspace files and write a recommendation resolving a complex billing dispute. Responses are scored on twelve criteria covering the correctness of the decision, reconciliation of the financial details, and adherence to policy. A pilot version was administered to the partner's subject-matter experts, whose feedback on clarity, accuracy, and relevance to the role was incorporated, including an extension of the time limit, and the experts reviewed the model answer. The partner engaged its legal function on the deterministic scoring method, seeking review under its standard assessment governance.

Participants and procedure. The assessment was administered to applicants in a live hiring process in summer 2026. Ninety-six applicants completed it, and the partner made 29 hires from this first round. Each response was scored by the engine and rated by a trained subject-matter expert using a behaviorally anchored rating form. The form provided criterion-level ratings corresponding to the twelve scored criteria and an overall rating on a five-point scale, on which the partner treats ratings of 3 or higher as passing. The overall rating is the comparison reported here.

© 2026 Codility. All rights reserved.

Analysis. For reporting, the partner grouped engine scores into five equal-width bands (0–20%, 21–40%, 41–60%, 61–80%, 81–100%) and cross-tabulated them against the expert's overall ratings. The bands were a convenience for display and were not designed to correspond to the expert's scale. Mean engine scores were computed at each rating level. Rank-order association was estimated with Spearman's ρ and Kendall's τb from the cross-tabulation, and classification agreement with the partner's pass threshold was examined at two engine cut points. Quadratic weighted kappa (QWK) was computed on the bands for completeness.

8.3 Results

Mean engine scores increased monotonically across the expert's rating levels (Table 15, Figure 1): from 37% for responses rated 1 to 60%, 76%, 83%, and 87% for responses rated 2 through 5. The rank-order association between the expert's rating and the engine band was substantial (ρ = .72; τb = .63). Against the partner's pass threshold of 3, the engine's classification matched the expert's for 79% of candidates at an engine cut of either 61% or 81% (κ = .56 and .59, respectively). These cuts were examined after the fact and are reported for illustration, not as recommended operating points. On the uncalibrated bands, the engine placed responses approximately one band higher than the expert on average (mean band 3.96 vs. 2.93). Exact band agreement was correspondingly low (22%), agreement within one band was 67%, and QWK was .50. The full cross-tabulation is reproduced in Appendix C.

Table 15. Study 3: Mean engine score by partner expert rating (N = 96)

Expert overall rating	Candidates	Mean engine score
1 (lowest)	12	37%
2	28	60%
3	24	76%
4	19	83%
5 (highest)	13	87%

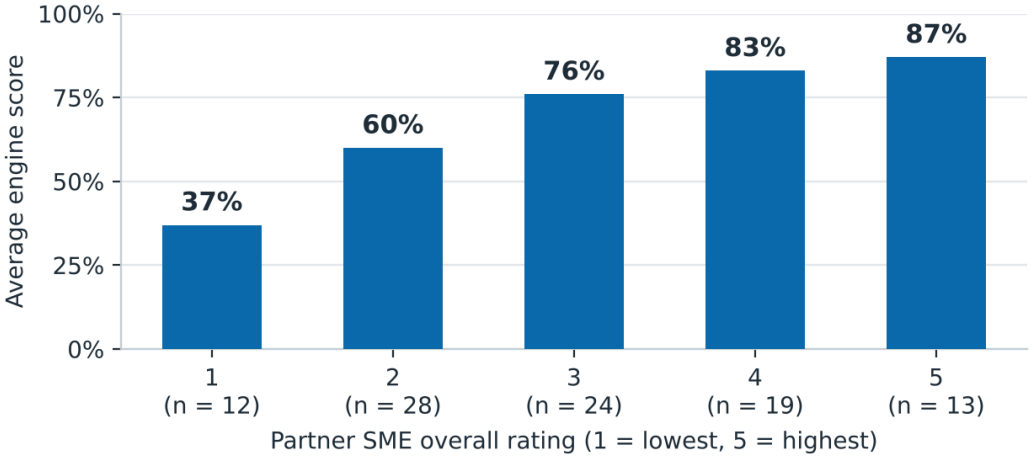


Figure 1. Study 3: mean engine score at each level of the partner expert's overall rating (96 applicants).

The largest group of discrepant cases consisted of responses the expert rated 2 that the engine placed above 60%: 14 of the 28 responses rated 2. In the partner's review of the results, its analysts reported that the expert's overall rating also took account of the professionalism of the written response, including grammar, spelling, and tone, which the twelve criteria do not score. The partner noted that written communication is assessed separately in its interview, so whether that dimension should be added to the

© 2026 Codility. All rights reserved.

assessment is an open design question rather than a scoring error. The expert's criterion-level ratings, which correspond directly to the scored criteria and would therefore be expected to agree more closely with the engine, were collected but had not been analyzed at the time of writing.

8.4 Observations reported by the partner's raters

The partner summarized its raters' observations in a debrief with Codility. Raters described the task as representative both of the work analysts perform and of the direction in which the role is moving, in which AI is used to extract insight from several sources at once, and therefore as a realistic preview of the job. They noted that the multiple-option structure of the scenario provided information about whether candidates read carefully and followed complex instructions. They found the three-panel interface easy to navigate. They reported that the assistant performed reliably, summarized case details accurately, and calculated credits and overcharges correctly. The rater also recommended three revisions to the scoring key. A criterion that credited either of two distinct actions should be split into two criteria. A single fact that was credited by three separate criteria should be consolidated so that it contributes one point. And criteria for grammar and spelling should be considered. In discussion, the partner's industrial-organizational psychology lead noted that this dimension could instead continue to be assessed in the interview.

8.5 Discussion and limitations

The monotonic relationship between engine scores and expert ratings, and the magnitude of the rank-order association, constitute convergent evidence for the scoring inference on authentic applicant responses, extending the constructed-response evidence of Study 1 to a different task and population. The narrowing of the difference in mean engine scores between expert ratings of 4 and 5 (83% vs. 87%) suggests that the engine discriminates less finely than the expert at the top of the range on this task. One possible explanation is the ceiling of a twelve-criterion key, which leaves limited room to distinguish strong from very strong responses. The offset between the engine bands and the expert's scale is not interpretable as a scoring bias, because the bands were not constructed to correspond to the scale; it does mean that any operational cut score for this task must be set on the engine's own metric rather than transferred from the expert's.

Four limitations qualify the results. First, a single expert rated all responses, so no human–human baseline is available. The equivalence criterion specified in the original plan, human–machine QWK of at least .70 and within .05 of the human–human value (Williamson et al., 2012), therefore could not be evaluated. Second, the study concerns one custom task. Third, the bands were defined after the fact. Fourth, because the study was embedded in live hiring, the expert's ratings may have informed hiring decisions, so they are best regarded as an alternative scoring of the same responses rather than as an independent criterion of job performance. The original plan also included an incumbent benchmark, which was not conducted because the assessment entered live hiring first. Several further analyses are in progress or planned. A second hiring round is under way with the task unchanged, after which both rounds will be rescored with the revised key so that the two versions can be compared on the same responses. The criterion-level ratings will be analyzed. Because the partner holds interview outcomes for these candidates, a first criterion-related analysis is also possible.

© 2026 Codility. All rights reserved.

9. Reliability and Generalizability

Reliability concerns the consistency of scores across the conditions under which they might have been obtained. Because scoring is deterministic, the substantive question is whether a candidate's score would have been similar on a different task of the same kind, which generalizability theory frames as the contribution of person-by-task interaction to score variance (Brennan, 2001). Research on performance assessment has consistently found that this source of variance is large, so that several tasks are needed to estimate a person's proficiency dependably (Shavelson et al., 1993). This section reports the evidence currently available on that question, which comes from a single pair of tasks completed by the same candidates, and describes the operational score distributions. The cross-task correlation indicates that the two tasks share variance consistent with a common capability, but its magnitude implies that single-task scores have limited precision and that multi-task batteries are needed for selection decisions.

9.1 Sources of measurement error

Because scoring is deterministic, scorer test–retest reliability is 1.0 and is not informative. Three sources of error remain. The first is the sample of tasks, which reflects the degree to which proficiency is task-specific. The second is occasion, which reflects variation in a candidate's performance from one administration to another. The third is the internal consistency of the criteria within a task. The data used in this report allow the first to be examined for one task pair. Test–retest data have not been collected. The criterion-level results required to estimate internal consistency were not part of the operational data export used here, and their analysis is part of the research agenda.

9.2 Correlation between tasks

In the operational data described in Section 10, 716 sessions came from a two-task battery in which candidates completed both AI Adoption, a decision briefing, and Budget Audit, an AI output audit. The two scores correlated r = .41 (Spearman ρ = .40). Restricting the analysis to external candidates, excluding internal test versions, yielded r = .41 (n = 572). The tasks differ in format, scenario, and evidence base, so the correlation reflects shared variance across substantially different task content. A correlation of this magnitude is consistent with the substantial task-specific variance documented for performance assessments (Shavelson et al., 1993), and it implies that a single task provides only a partial estimate of a candidate's proficiency.

9.3 Projected consistency of multi-task composites

If the two tasks are treated as alternate measures of a common capability, the Spearman–Brown formula projects the consistency of a composite of k tasks from the inter-task correlation. With r = .405, the projected consistency is .58 for two tasks, .67 for three, and .73 for four. These projections assume that additional tasks would be as related to one another as the two observed, which may not hold across formats or role families, so they should be read as approximate. They nonetheless indicate the order of magnitude of the gain from adding tasks, and they are the basis for the recommendation in Section 14 that selection decisions rest on batteries of at least two tasks, and preferably three or more where stakes are high.

© 2026 Codility. All rights reserved.

9.4 Operational score distributions

Table 16 summarizes the distributions of task scores in the July–October 2026 operational data. Scores span the full range, mean task scores lie between 38% and 46%, and there is no indication of a ceiling effect. A substantial share of scored administrations received zero, from 11% for AI Adoption to over 30% for the less frequently administered tasks; the export does not indicate whether these were blank submissions or attempted responses that met no criteria. In the combined score, tasks that were not scored are counted as zero, as the platform counts them. This convention affects 48 of the 1,006 sessions, 12 of which contain no scored AI for Business task at all, and the combined-score row of Table 16 includes these substituted values. A substituted zero records the absence of a score, not an observed performance of zero. Separating non-attempts from attempted responses is a prerequisite for estimating reliability without their influence and for constructing local benchmarks.

Table 16. Operational score distributions, July to October 2026

Task	Format	Scored administrations	Mean	SD	Median	Scored zero
AI Adoption	Decision briefing	770	46.3	22.8	50	11%
Budget Audit	AI output audit	716	41.3	31.9	50	25%
Office Relocation	Research drilldown	136	42.8	37.4	40	32%
Phantom Lift	Decision briefing	76	38.0	31.6	42	32%
Supplier Dispute	Prompt engineering	23	41.7	39.4	30	35%
Combined score (sessions)		**1,006**	**42.0**	**27.1**	**41.5**	**14%**

© 2026 Codility. All rights reserved.

10. Subgroup Differences and Adverse Impact (Study 5)

The *Uniform Guidelines* (1978) require employers to monitor whether a selection procedure produces substantially different selection rates across demographic groups, and the *Standards* (AERA, APA, & NCME, 2014) call for evidence on subgroup differences as part of the case for a procedure's use. For a new assessment category, such evidence can only accumulate as operational use grows, and early analyses carry correspondingly greater uncertainty. Study 5 is the first such analysis for AI for Business. It examined standardized mean differences and selection rates by gender and by race or ethnicity in 1,006 operational sessions, 330 of which included self-reported demographic information. Differences between women and men were negligible, and no analysis indicated adverse impact against women. Standardized mean differences among racial and ethnic groups were small. At the more selective hypothetical cut scores, however, the selection rates of White and of Hispanic or Latino candidates fell below four-fifths of the rate for Asian candidates in several analyses. These results are reported in full below, together with characteristics of the sample that limit how far they can be generalized.

10.1 Analytic framework

Adverse impact was evaluated in three complementary ways. Practical significance was assessed with the impact ratio, computed in two forms. For the pairwise comparisons in Table 19, it is the lower selection rate of the pair divided by the higher. For the group-level summary in Table 18, following the *Uniform Guidelines*, it is each group's selection rate divided by that of a reference group, defined as the group with the highest selection rate among groups with at least 30 members; groups with fewer than 30 members could not serve as the reference, so their ratios can exceed 1.00. Both forms were evaluated against the four-fifths guideline, supplemented by the one-person flip rule, which asks whether an impact ratio below .80 would cross .80 if one additional member of the lower-rate group passed. Statistical significance was assessed with two-proportion Z tests and Fisher's exact tests (Morris & Lobsenz, 2000). Mantel–Haenszel tests stratified by test version guarded against artifacts of aggregating across batteries (Mantel & Haenszel, 1959). Standardized mean differences were estimated as Hedges' g with 95% confidence intervals and tested with Welch's t, with least-squares models including test fixed effects as a robustness check. Because no operational cut score had been set by any employer in the sample, selection rates were evaluated at four hypothetical cut scores (40, 50, 60, and 70), corresponding to overall selection ratios from .54 to .18.

10.2 Method

Data. The data comprise 1,006 Codility Screen sessions from July to October 2026 that included AI for Business tasks. In 48 of these sessions at least one AI for Business task was not scored, and in 12 no AI for Business task was scored. Following the platform's convention, the combined score counts an unscored task as zero. A substituted zero is not an observed zero, and results for the combined score should be read with that convention in mind. Among the sessions with demographic information, six include an unscored Budget Audit task, which the analysis records as zero in both the Budget Audit and the combined scores; excluding these six sessions changes no standardized difference by more than 0.03 and moves no impact ratio across 0.80. These include 161 sessions from internal test versions, which the sensitivity analyses exclude. Demographic information was available only for candidates who completed the optional post-test survey and reported their gender or race and ethnicity using EEO-1 categories: 330 sessions (33%). Survey

© 2026 Codility. All rights reserved.

respondents scored slightly higher than non-respondents (d = 0.15, p = .013). Most sessions, and 94% of sessions with demographic information, came from a two-task battery (AI Adoption and Budget Audit) administered by a global enterprise software company to a multinational applicant pool.

Measures and groups. The outcomes were the combined AI for Business score (the mean of the tasks taken) and the two task scores. The groups analyzed and the minimum standardized differences detectable with 80% power at α = .05 are shown in Table 17. Groups with fewer than 30 members were reported descriptively only. All analyses were repeated with internal test versions excluded, and all 2,472 statistics computed in the analysis workbook were reproduced independently in Python, with complete agreement.

Table 17. Study 5 analysis groups and minimum detectable differences

Group	n	Minimum detectable difference (d)
Women / men	123 / 193	0.32
Asian / White	109 / 125	0.37
Hispanic or Latino (vs. White or Asian)	47	0.48–0.49
Black or African American; Other (combined)	9; 9	Descriptive only

10.3 Summary of results

Table 18 summarizes the results for the combined score in simplified form; the subsections that follow report them in detail. Two statistics are shown. Cohen's d expresses the difference between two groups' average scores in standard-deviation units; by convention, values below 0.20 are negligible, 0.20 to 0.49 small, 0.50 to 0.79 medium, and 0.80 or more large. The impact ratio divides a group's pass rate by that of a reference group, the group with the highest pass rate among groups with at least 30 candidates; under the four-fifths guideline of the *Uniform Guidelines*, a ratio below 0.80 is generally treated as evidence of adverse impact that warrants review. Because no employer in the sample had set a cut score, impact ratios are shown at four hypothetical cut scores.

In summary, average scores differed little between any two groups of adequate size (|d| ≤ 0.30). Women and men passed at similar rates; the only ratio below 0.80 (0.797, for men at a cut score of 70) would be reversed if one additional man had passed. At cut scores of 40 and 50, every group of adequate size passed at no less than four-fifths of the highest rate. At cut scores of 60 and 70, White and Hispanic or Latino candidates passed at less than four-fifths of the rate of Asian candidates, who had the highest pass rate among groups of at least 30 at every cut.

Table 18. Study 5 at a glance: score differences and impact ratios on the combined score

Group	n	Average score	Difference in average score (Cohen's d)	Impact ratio, cut 40	Impact ratio, cut 50	Impact ratio, cut 60	Impact ratio, cut 70
Women	123	45.9	+0.07 vs. men (negligible)	0.97	0.97	1.00	1.00
Men	193	44.2	Comparison group	1.00	1.00	0.84	**0.797**
Asian	109	47.6	+0.14 vs. White (negligible)	1.00	1.00	1.00	1.00
White	125	44.4	Comparison group	0.90	0.86	**0.71**	**0.52**
Hispanic or Latino	47	40.7	−0.17 vs. White (negligible); −0.30 vs. Asian (small)	0.85	0.81	**0.76**	**0.43**
Black or African American†	9	38.3	−0.28 vs. White (small)	**0.77**	0.85	**0.56**	0.90

© 2026 Codility. All rights reserved.

Other (combined)†	9	47.2	+0.13 vs. White (negligible)	1.15	1.06	0.85	**0.45**

Note. Combined AI for Business score; 330 survey respondents who reported gender or race and ethnicity. Impact ratio = the group's pass rate divided by the pass rate of the reference group in the same comparison (gender, or race and ethnicity) at that cut score. The reference group, marked 1.00, is the group with the highest pass rate among those with at least 30 candidates; smaller groups were not eligible, which is why some of their ratios exceed 1.00. Six respondents' combined scores include a substituted zero for an unscored task (Section 10.2). Ratios below 0.80 are shown in bold. Positive d indicates a higher average score than the comparison group. The values of d agree with the Hedges' g values in Table 19 to two decimal places. †Fewer than 30 candidates; shown for completeness and not interpreted.

10.4 Results by gender

Differences between women and men were negligible on the combined score (g = 0.07, 95% CI [−0.15, 0.30]), on AI Adoption (g = 0.15), and on Budget Audit (g = −0.04), and none was statistically significant. Impact ratios met the four-fifths guideline in 10 of 12 analyses. Both exceptions occurred at the highest cut score and involved lower selection rates for men. On the combined score the impact ratio was .797, which a one-person change would reverse. On AI Adoption it was .52 (Fisher p = .084; Mantel–Haenszel p = .037). Where women had the lower selection rate, the smallest impact ratio was .89. Excluding internal test versions did not change the pattern.

10.5 Results by race and ethnicity

Standardized mean differences were small (Table 19, Figure 2). Asian candidates scored higher than White candidates by g = 0.14 to 0.26 across the three outcomes, none statistically significant at α = .05. Hispanic or Latino and White candidates did not differ meaningfully (|g| ≤ 0.17). Hispanic or Latino candidates scored lower than Asian candidates by g = −0.18 to −0.37, a difference that was statistically significant on AI Adoption only (g = −0.37, p = .040).

At cut scores of 40 and 50, every comparison between groups of adequate size met the four-fifths guideline or fell within one candidate of it. At the more selective cut scores of 60 and 70, Asian candidates had the highest selection rates. The impact ratio for White candidates fell below .80 in 6 of 12 analyses, 4 of which were significant by Fisher's exact test. The selection rate of Hispanic or Latino candidates fell below four-fifths of that of Asian candidates in 6 analyses, none statistically significant and 3 reversible by a one-person change. Of the 84 selection-rate tests in the primary sample, 7 were significant at α = .05, compared with about 4 expected by chance. All 7 reflected lower selection rates for White candidates relative to Asian candidates or to all non-White candidates combined.

Table 19. Study 5: Standardized differences and selection-rate analyses (all respondents)

Comparison (focal vs. reference)	g, combined [95% CI]	g, AI Adoption	g, Budget Audit	Impact ratio < .80 (of 12)	Fisher p < .05
Women vs. men	0.07 [−0.15, 0.30]	0.15	−0.04	2 (men lower)	0
Asian vs. White	0.14 [−0.11, 0.40]	0.26	0.21	6 (White lower)	4
Hispanic or Latino vs. White	−0.17 [−0.50, 0.17]	−0.15	0.02	2 (one each)	0
Hispanic or Latino vs. Asian	−0.30 [−0.64, 0.05]	−0.37*	−0.18	6 (Hispanic or Latino lower)	0

© 2026 Codility. All rights reserved.

Note. Positive values of g indicate higher scores for the focal group. *Welch p < .05. Impact-ratio counts cover three outcomes at each of four hypothetical cut scores.

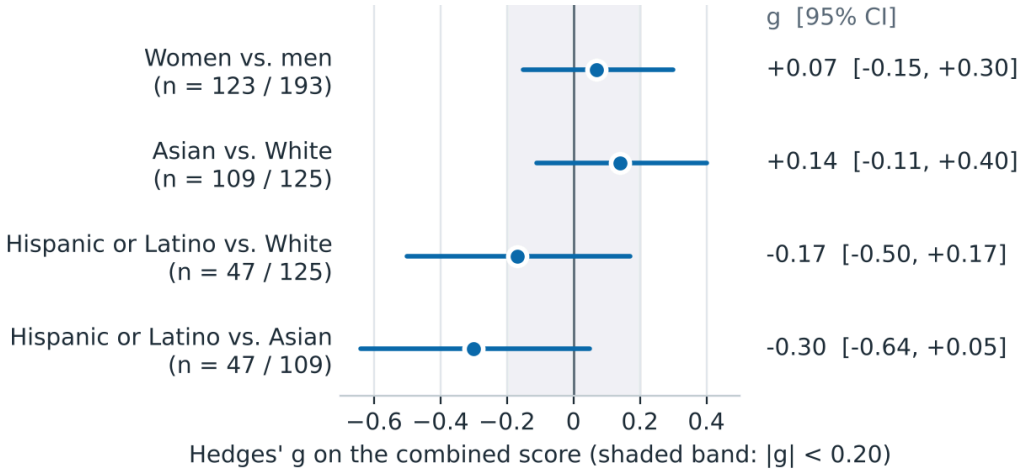


Figure 2. Study 5: standardized mean differences on the combined score with 95% confidence intervals.

10.6 Discussion and limitations

The results for gender provide no indication that the assessment disadvantages women; where the impact ratio fell below .80, the lower rate was for men, and only at the most selective cut. The results for race and ethnicity are more complex. Mean differences were small by conventional standards. The selection-rate differences at selective cuts arise largely from the comparatively high selection rates of Asian candidates, the highest-rate group; White and Hispanic or Latino candidates did not differ meaningfully from each other. Under the *Uniform Guidelines*, however, impact ratios are evaluated relative to the highest-rate group of adequate size regardless of which group that is, and the results at cuts of 60 and 70 would warrant attention if an employer adopted a cut in that range.

Several characteristics of the sample limit interpretation. The cut scores are hypothetical, and adverse impact is properly evaluated at the cut score an employer actually uses. Demographic data are available only for the third of candidates who completed the survey, and those candidates scored modestly higher than non-respondents. Nearly all of the demographic sample came from one employer's multinational applicant pool, to which U.S. EEO-1 categories were applied without information on candidates' country of residence. Differences between groups may therefore partly reflect the regional composition of the applicant pool, including differences in English proficiency, rather than differences attributable to the assessment. Age and disability status were not collected. The analyses proposed in Section 14 follow from these limitations: analysis by country, analysis by age (40 and over), and analysis of employer-specific operational cut scores as they are set. Employers deploying the assessment at volume should monitor selection rates in their own applicant flow from the outset, and Codility's Assessment Science team can support the design of that monitoring.

© 2026 Codility. All rights reserved.

11. Candidate Reactions

Applicants' perceptions of a selection procedure bear on its acceptability and are related to their attitudes toward the hiring organization and their intentions to remain in the process (Gilliland, 1993; Hausknecht et al., 2004). Work samples are generally among the procedures applicants view most favorably, but an assessment that requires candidates to work with an AI assistant is unfamiliar, and its reception cannot be assumed. This section reports reactions collected through the post-test survey from candidates in the operational data described in Section 10. Ratings were most favorable for the AI assistant and the testing environment and least favorable for items concerning the job-relatedness of high scores. Perceptions of fairness did not differ between women and men. Candidates' written comments, summarized at the end of the section, describe both what they valued and where the experience fell short.

11.1 The survey instrument

The post-test survey was redesigned in July 2026 for candidates taking business tasks. The redesign reduced the instrument from 22 to 16 questions with conditional logic and removed all references to programming. It adopted balanced four-point labels for rating items and fully labeled five-point scales for agreement items. Demographic questions were moved to EEO-1 and U.S. Census education categories, and a module on the AI assistant was added. Because response labels changed, results from the redesigned survey form a new baseline and are not directly comparable with earlier results from the coding survey. Responses are voluntary, and the response rate in the operational data was 33%, so the results may not represent all candidates.

11.2 Results

The percentage of respondents giving a favorable response, defined as good or very good on rating items and agree or strongly agree on agreement items, is shown in Figure 3. The AI assistant (81%) and the testing environment (79%) received the most favorable ratings, and 75% of respondents agreed that the task description was clear. Sixty-nine percent rated the test favorably as an evaluation of their skills, and 64% agreed that the task reflected real work. Fifty-seven percent agreed that the content seemed fair (28% neutral, 15% unfavorable), 56% that the assessment gave them an opportunity to show their skills (20% neutral, 24% unfavorable), and 51% that the task evaluated their skills well. The least favorable item asked whether a person who scored well would be a good candidate for the job. On that item, 37% agreed, 42% chose the neutral midpoint, and 21% disagreed, a pattern that may reflect the difficulty applicants have in judging the job-relatedness of a novel assessment from the candidate's position.

© 2026 Codility. All rights reserved.



Figure 3. Candidate reactions: share of respondents responding good or very good, or agree or strongly agree, July to October 2026.

11.3 Differences between women and men

Women and men did not differ in perceived opportunity to perform, content fairness, job-relatedness, measurement quality, task realism, or testing experience ($|d| \leq 0.24$; all Holm-adjusted $p \geq .90$). A three-item fairness composite, whose items reflect the procedural-justice dimensions described by Gilliland (1993), was internally consistent ($\alpha = .85$) and likewise showed no difference. Women estimated their own scores lower than men did (57.8 vs. 65.4; $d = -0.36$; Holm-adjusted $p = .035$), although their actual scores did not differ (46.1 vs. 44.5). Both groups overestimated their scores, men by more (mean overestimate of 21.3 vs. 11.6 points). The discrepancy between estimated and actual scores is relevant to the design of the feedback candidates receive.

11.4 Candidate comments

The survey included four optional open-ended questions, and 274 of the 396 respondents answered at least one. The comments below were selected to illustrate recurring themes and are not a representative sample; they are reproduced as written, with spelling corrected. Favorable comments most often concerned the realism of the tasks, the clarity of the environment, and the responsiveness of the assistant. Others, including comments from candidates without a technical background, valued the practice environment and the experience of applying AI to an unfamiliar business problem.

> *"I like this it is measuring AI fluency, using AI as a thought partner and not just for answers, and it requires critical thinking."* — Candidate, September 2026

> *"Realistic scenarios: that mirror actual analyst work … Critical thinking: required to rule out three options, not just pick one."* — Candidate, September 2026

> *"As someone not so familiar with IT and more exposed on Finance, I really appreciate the test on how to use the AI tools."* — Candidate, August 2026

> *"Having this environment available to practice was a great option. By the end of the demo test, I felt much more confident about completing my assessment."* — Candidate, July 2026

© 2026 Codility. All rights reserved.

The most frequent suggestions concerned the time limit and the ability to copy and paste; 44 respondents mentioned copying or pasting. The copy restriction is a deliberate content-protection measure (Section 12.2), and a recalibration of task time limits from observed completion times has been proposed. Candidate comments are summarized for the product and content teams in regular read-outs, and some suggestions have already led to changes. Until August 2026, the attached materials could be reached only by asking the assistant, and candidates asked to be able to read them directly:

> *"Clarify if there are documents we can open or if we have to ask the AI all of the extra details to analyze the business case."* — Candidate, July 2026

Client organizations had made the same request. In August 2026, a file viewer was added to the task panel. Candidates can now open each attached file in its own tab and read it beside the assistant and the response editor, with data tables shown as read-only spreadsheets and documents as formatted text; the files cannot be downloaded, copied, or printed (Section 12.2). After the release, one candidate noted that the attachments could be viewed "in a separate tab without closing any existing windows." Comments about where the task data could be found continued after the release, however, which suggests that the task instructions should also state which materials are attached and what the assistant can see.

© 2026 Codility. All rights reserved.

12. Language, Test Security, and Assistance from External AI

Three conditions of administration can affect the meaning of a score independently of the construct: the language in which the assessment is delivered, the security of its content, and the availability to candidates of AI tools outside the task environment. All three were identified as potential sources of construct-irrelevant variance in Section 2.4. The available evidence on language effects is limited, but it does not indicate a disadvantage for candidates in two large markets where English is not the primary language. The data-access design protects research drilldowns and AI output audits well against answers produced by an external AI from the task description alone. Decision-briefing and prompt-engineering tasks are exposed, the short-form tiers have not yet been tested in this way, and a hardening program has been defined that covers every live task.

12.1 Language

All tasks are delivered in English. The design limits the role of language proficiency in several ways. Task descriptions have been shortened, the criteria evaluate principally the substance of the conclusion, and grammar and spelling are not scored. Evidence on whether English proficiency nonetheless affects scores is limited to an operational summary prepared in August 2026 for a global enterprise software company. In that summary, candidates in Japan (mean normalized score 74.3) and China (70.2) scored at or above candidates in the rest of the world (68.3). The summary used a normalized metric that differs from the percentage scores reported in Section 9 and did not report group sizes or variability, so it cannot be treated as a formal analysis, and a country-level analysis is proposed. A review of the reading level of business tasks is also proposed, applying the CEFR-based method Codility previously used for its coding content (Codility, 2026a).

A related question concerns standardization. When asked, the embedded assistant will summarize task content in another language, and a customer's test of this behavior showed that it provides a summary rather than a full translation. Candidates who request such help therefore experience somewhat different conditions from those who do not. Whether and how language support should be offered is accordingly treated as a condition of administration to be standardized, and a customer request for localized delivery is under review.

12.2 Test security and the data-access design

The principal security feature of the design is that the information needed to answer each task is available only inside the environment, through the in-task assistant and the read-only file viewer, so that a description copied outside the environment does not carry the evidence base with it. The assistant's behavior specification controls what it discloses and when, and the complete conversation is attached to the candidate report, giving reviewers a record of the candidate's process. Platform safeguards documented in the *Codility Technical Manual*, including paste detection, identity verification, and optional proctoring, also apply. During the preview period, paste detection in the text editor initially failed to register pasted content; the defect has since been corrected. In August 2026, copying from the assistant's conversation and from the file viewer was also blocked, and attached files can no longer be downloaded or printed, which makes task content harder to extract from the environment.

© 2026 Codility. All rights reserved.

12.3 Assistance from external AI

Study 2 showed that the protection afforded by the data-access design depends on format (Table 12). Research drilldowns and AI output audits were rarely solvable from the description alone. Decision briefings and prompt-engineering tasks frequently were, because their descriptions carry information about the answer: the list of options in the first case, and the specification of the prompt's purpose in the second. Under a hardening program scheduled to begin in October 2026, every live task will be tested by submitting, unchanged, the answer an external model produces from the task description, and each task must score below an agreed threshold on that test. The design changes under consideration are: holding decisive facts in the evidence base instead of the description; designing the assistant's first answer to be fluent but incomplete or wrong, as is already done in the inbox tier; requiring facts that surface only from a well-formed query of the sources; and introducing decision checkpoints within a simulation, a pilot of which is scheduled. Until a task has been hardened, proctoring is recommended for the two exposed formats. Reviewers also observed that an external model can suggest useful questions to put to the in-task assistant. Because credit depends on reaching and supporting the correct conclusion from the evidence, a candidate who uses such questions effectively may still need to exercise part of the construct, but the degree to which this inflates scores has not been estimated.

© 2026 Codility. All rights reserved.

13. Measuring the AI Interaction (Study 6)

Everything scored at present is the written work product. As Section 5.3 showed, the written product reflects some components of AI collaboration, particularly validation, more directly than others, such as delegation and iteration, which are properties of the interaction itself. Measuring the interaction would therefore reduce the construct underrepresentation identified in Section 2.4. Doing so defensibly is, however, a harder problem than scoring a constructed response against a defined answer space, because there is no single correct interaction. The research program described here is sequenced to address that difficulty in stages: descriptive measures first, then structured summaries that support reviewers without producing a score, and finally scores, each step to be validated before the next is used in decisions. Study 6, a program of interim studies conducted on engineering transcripts, found that language-model judgments of transcripts can be made highly repeatable but that repeatability did not predict agreement with expert judgment. In the first comparison with an expert's ratings, a holistic judgment agreed moderately with the expert, whereas methods built from behavioral indicators, including the original indicator bank, which had been the most stable, agreed poorly. No transcript-derived score is used in decisions.

13.1 Planned capabilities

Table 20 lists the capabilities in development or planned, and Table 21 defines the four descriptive metrics of the first release. The metrics are named as literal measurements, distinct from the names of the competencies reported alongside them, so that observed behavior is not confused with inferred skill. Three of the four correspond to subskills of AI Collaboration and the fourth to a subskill of AI Engineering. Each is a proxy for behavior rather than for the quality of judgment behind it, which is why the first release makes no skill-level claims. Absolute values are always shown. A comparative band (lower than, typical of, or higher than the cohort) is shown only within the same assessment and environment, and only once the cohort reaches a minimum size. The framing parallels the AI Fluency framework of Dakan and Feller (2025).

Table 20. AI interaction measurement: planned capabilities

Capability	Nature of the output	Status (October 2026)
Verification sub-score	Credit on the final answer for identifying planted errors (e.g., two of three), derived from existing criteria	Scheduled (from October 2026)
Version 1: interaction metrics	Descriptive measures reported beside the task outcome; no skill claims	In development
AI fluency profile	Banded results for each skill, aggregated across tasks	Scheduled (from October 2026)
Version 2: transcript summaries	Rubric-anchored summaries of the candidate's process to aid review; not scored	Prototype (Study 6)
Version 3: transcript scoring	Per-skill scores, with the transcript evidence for each shown in the report	Research and development

© 2026 Codility. All rights reserved.

Table 21. Version 1 interaction metrics

Metric	Definition	Related ESM subskill
AI-Generated Share	Proportion of the submitted solution that originated with the AI rather than the candidate	AI Task Delegation
Prompts Sent	Number of times the candidate instructed or redirected the AI	AI Iteration
Test Runs	Number of times the candidate tested or verified output after the AI produced it	AI Output Validation
AI Usage	AI consumption during the session, as a share of the assessment's AI budget	AI Cost Optimization

13.2 Study 6: Transcript-evaluation methods

Purpose and data. Study 6 comprises a sequence of studies, conducted between July and October 2026, of the methods on which transcript summaries and, eventually, transcript scores would rest (Codility, 2026e, 2026f, 2026h, 2026i, 2026j, 2026k). All of them used transcripts of AI-assisted software engineering rather than AI for Business sessions, and all except the study of displayed evidence (Codility, 2026k), which covered five ESM AI skills, target a single construct, context engineering (ESM subskill 10.3), defined as structuring requests so that the AI has the objective, context, constraints, specifications, and success criteria it needs. The first two phases used an external corpus of naturalistic developer–assistant sessions (SWE-chat), each accompanied by a session-success score generated by a language model from a rubric. Of 4,513 eligible sessions, 1,868 fell within a range of 500 to 5,000 characters written by the user, chosen to approximate the volume of interaction in a 15- to 60-minute assessment. Two hundred were sampled for the first phase, 50 from each of four bands of the success score, and the sample was extended to 300 for the second. The later phases used transcripts of Codility candidates who completed engineering tasks while using AI coding assistants.

Method. Two architectures were compared. In the holistic architecture, a language model read the full transcript and assigned a single 1–5 rating of the user's context engineering. In the granular architecture, consistent with the test-case approach used in written-product scoring, one model call selected the user messages containing scorable evidence, a second judged each of twelve behavioral indicators, spanning novice to expert proficiency, as passed or failed, and a weighted aggregation converted the outcomes into a score between 0 and 1 in which passing a high-level indicator and failing a low-level indicator carry the most information. For comparison with the holistic ratings, granular scores were also converted to five bands of equal width. From the second phase onward, a ten-indicator bank written by subject-matter experts was scored alongside the original twelve-indicator bank, which had been drafted with the assistance of a language model. Each transcript was scored three times under each method, and consistency was examined for the overall classification, the indicator judgments, and the evidence selected.

Results on the external corpus. Both architectures were highly repeatable in the first phase (Table 22). On the common five-point scale, ICC(3,1) was .910 for the holistic architecture and .912 for the granular architecture, and on its original continuous scale the granular score reached .939. The granular architecture produced more stable classifications: 21.0% of sessions changed band across the three runs, compared with 30.8% under the holistic architecture, a reduction of about one-third (Figure 4). Its component judgments were also stable. Across 2,400 indicator-level decisions, pairwise agreement between runs ranged from .956 to .964 (κ = .898 to .916), and the same evidence set was selected in all three runs for 72% of sessions. Two qualifications emerged from the same data. Granular scores were concentrated in

© 2026 Codility. All rights reserved.

the fourth band (118 to 123 of 200 sessions per run), which may account for part of the granular architecture's advantage in classification stability, and the scores of both architectures were positively associated with the amount the user wrote. In an exploratory analysis of criterion-related evidence, neither architecture was related to the continuous session-success score (r = −.15 for the holistic and −.00 for the granular architecture). In a second sample of 200 sessions balanced for fully resolved outcomes, neither was related to whether a session was fully resolved (point-biserial r = −.04 to .06; AUC = .46 to .54). Because the success score was itself generated by a language model, and because success in a naturalistic session depends on task difficulty, the assistant's performance, and many factors outside the user's requests, these null results are difficult to interpret. They nonetheless provide no criterion-related support for the construct as measured.

The second phase, with 300 sessions, replicated these findings (Codility, 2026h). Run-to-run reliability was .915 for the original indicator bank, .899 for the holistic rating, and .864 for the expert-written bank. Agreement on individual indicators was nearly identical for the two banks (κ = .867 and .862), and about 40% of the difference between their composite reliabilities was attributable to the expert bank's smaller number of indicators. Relationships with session success were again null (r = −.09 to .02; AUC = .47 to .55), and scores again tracked the amount the user wrote (r = .41 to .45). The holistic method declined to rate six of the shorter sessions in at least one run, so its reliability estimate rests on 294 sessions; the indicator methods, by contrast, always return a score.

Replication with Codility candidates. In September 2026 the procedures were applied to Codility candidates who completed engineering tasks while using AI coding assistants, first to 92 candidates and then to 233 candidates on a single task (Codility, 2026f, 2026i). Run-to-run consistency was again high (ICC(3,1) = .850 to .949 in the first sample and .884 to .948 in the second), and the expert-written bank was again the least consistent; in the 92-candidate sample, the difference in the number of indicators explained only about 22% of the reliability gap between the two banks. Four findings qualify that result. Both indicator banks reached a ceiling on the single-task sample: 77% of candidates fell in the top band of the original bank, and most indicators were passed by 89% or more of candidates. Although the methods agreed in identifying weak candidates, they ordered the remaining candidates differently, with rank-order correlations between methods of only .43 to .59. Scores were more strongly related to the amount the candidate wrote than in the external corpus (r = .55 to .72 with the logarithm of characters written). And when a candidate used two assistants, combining the two transcripts credited the candidate's best moment rather than typical behavior under the indicator methods: for the eleven candidates in the 92-candidate sample who used both assistants substantively, the combined score fell at or near the higher of the two single-assistant scores.

Table 22. Study 6: Interim results on the external corpus (200 sessions × 3 runs)

Measure	Holistic	Granular (evidence + indicators)
Test–retest reliability, ICC(3,1), common 1–5 scale	.910	.912
Sessions changing classification across 3 runs	30.8%	21.0%
Indicator-level cross-run agreement (2,400 decisions)	—	.956–.964 (κ = .898–.916)
Identical evidence set selected across all 3 runs	—	72% of sessions
Test–retest reliability, ICC(3,1), original scale	.910 (1–5)	.939 (0–1)
Correlation with continuous session-success score	−.152	−.004
Point-biserial r with fully resolved session (separate balanced sample of 200; 3 runs)	.046 to .060	−.036 to −.021

© 2026 Codility. All rights reserved.

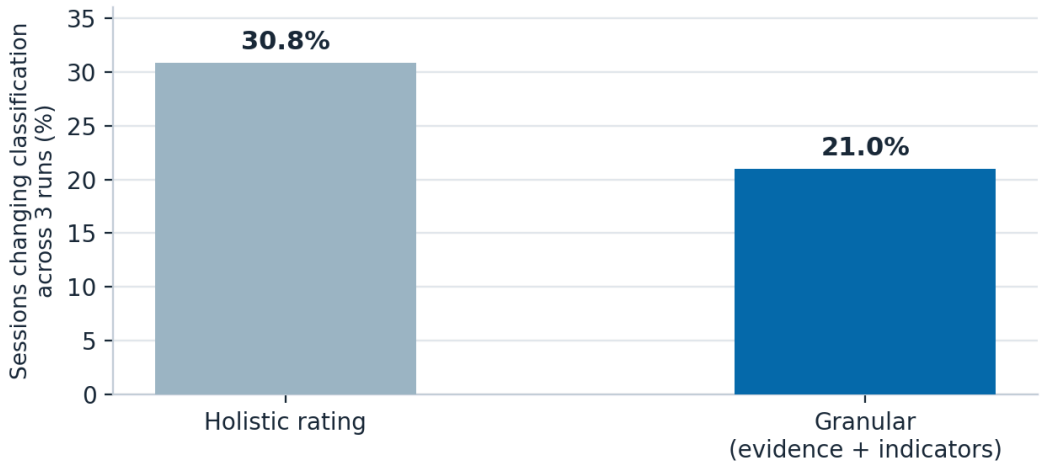


Figure 4. Study 6, first phase (200 external sessions, common 1–5 scale): percentage of sessions changing proficiency classification across three evaluation runs under each architecture (lower values indicate greater stability). Part of the difference may reflect the concentration of granular scores in one band.

Agreement with expert ratings. The first comparison with expert judgment was completed in October 2026 (Codility, 2026j). One subject-matter expert rated the context engineering of 70 of the 233 candidates holistically on the five ESM proficiency levels, and the 61 candidates with at least three turns of interaction formed the analysis sample. Before the comparison, three success criteria were fixed for a new method that derives a proficiency level from indicator evidence: absolute agreement with the expert of ICC(A,1) ≥ .55 on transcripts not used to select the method's placement rule, a mean difference from the expert within ±0.3 levels, and run-to-run agreement on the method's new indicators of κ ≥ .80. The method met none of them. Its agreement with the expert was ICC(A,1) = .11 (95% CI −.02 to .26; .06 in a second run), it placed candidates 0.62 levels below the expert on average, and run-to-run agreement on its new indicators was κ = .72. Table 23 compares it with the other methods on the same 61 transcripts. The holistic language-model rating agreed moderately with the expert (ICC(A,1) = .63), whereas every method built from behavioral indicators agreed poorly (.11 to .26), and the two indicator banks rated candidates more than a level above the expert on average. The analysis attributed the failure of the evidence-based level methods to the evidence they draw on: display indicators are set at a deliberately high bar and record a candidate's strongest moments, whereas a holistic rating judges typical behavior, and 37 of the 61 candidates produced no such evidence either way. The two indicator banks erred in the opposite direction, placing candidates more than a level above the expert. The expert's ratings also gave credit for restating the task to the assistant, which the scoring rules exclude by design, and tracked the amount candidates wrote in their own words (ρ = .41); these features explain part, but not all, of the gap. The holistic rating likewise credited restating the task (by 0.93 levels, compared with 0.55 for the expert, among the 16 candidates who did so), which may account for part of its agreement with the expert. The comparison rests on one expert and one fully specified task, so expert–expert agreement, which the study plan requires as a benchmark (Codility, 2026g), is not yet available.

Table 23. Study 6: Agreement with one expert's holistic ratings (61 candidates)

Method	ICC(A,1) [95% CI]	Rank correlation	Mean difference (method − expert)
Holistic language-model rating	.63 [.48, .75]	.64	+0.20
Original indicator bank (weighted pass rate)	.19 [.08, .31]	.43	+1.51
Expert-written indicator bank (weighted pass rate)	.26 [.10, .43]	.39	+1.10

© 2026 Codility. All rights reserved.

Highest level evidenced by display indicators	.17 [−.01, .34]	.21	−0.83
Indicator-based proficiency level (held out)	.11 [−.02, .26]	.19	−0.62

Reliability of displayed evidence. Separately, the step that selects transcript evidence for display to reviewers, on which the transcript summaries in Table 20 would draw, was evaluated for reproducibility across five ESM AI skills: context engineering, AI iteration, AI output validation, AI task delegation, and AI failure modes (Codility, 2026k). In two complete runs on the 233 candidates, verdicts differed on 0.5% to 1.3% of message-by-indicator checks, agreement at the level of candidate and indicator was κ = .83 to .88, and the decision whether to display a skill for a candidate was unchanged for 89.9% to 97.1% of candidates with at least three messages. Text that a candidate copied from the task description is masked before judgment, with no false masks in a negative control of 2,001 messages, and every pass must quote the candidate's own words. Whether the displayed evidence is correct has so far been checked only in internal audits of 30 items: after the check lists were tightened, 2 of 30 audited context-engineering passes were still judged wrong, and 4 of 30 audited passes still credited restated task requirements; a blind review of a 137-item set by two independent reviewers is in preparation.

Discussion and next phase. Taken together, the studies show that language-model judgments of transcripts can be made highly repeatable, and they show equally clearly that repeatability does not establish accuracy. The original indicator bank was the most stable method but agreed poorly with the expert, and none of the methods was related to the external corpus's measure of session success. The results do not support using transcript-derived scores in selection decisions, and none is used. They are more favorable to a narrower use, the display of quoted transcript evidence to human reviewers, subject to the independent review of its correctness now in preparation. For scoring, the analysis proposes a structured holistic judgment that places the candidate directly on the ESM level descriptions and cites the supporting evidence, which is intended to retain the agreement shown by the holistic rating while making the judgment auditable; it has not yet been validated against expert ratings. That design requires fresh expert ratings, a second expert to establish expert–expert agreement, and evaluation on more than one task. Extension from engineering transcripts to AI for Business transcripts, and from context engineering to the other ESM AI skills, would follow. No transcript-derived score will be used in a selection decision until that evidence is available.

© 2026 Codility. All rights reserved.

14. Intended Use, Limitations, and Research Agenda

The preceding sections assembled evidence for each inference in the validity argument outlined in Section 1.4. This section draws that evidence together. It states the uses the evidence currently supports, the uses that would require further or local evidence, and the limitations that apply to all of them. It closes with the research agenda through which the remaining questions will be addressed. In summary, the evidence supports using AI for Business scores to rank-order candidates within multi-task batteries, with human review of mid-range scores and with cut scores set locally. It does not yet support stand-alone high-stakes use of the short-form tasks, normative interpretations, or claims of criterion-related validity.

14.1 Uses supported by the current evidence

The evidence supports the use of AI for Business scores, based on the scored work product, as one component of structured screening and evaluation. The recommended model is rank-ordering within a battery. Because single-task scores have limited precision (Section 9), batteries should include at least two tasks, and three or more where decisions carry high stakes. Consistent with Codility's principle of human oversight and with the finding of Study 1 that agreement with experts is weakest for partially correct responses, mid-range scores should receive supplementary human review. Where tasks in the decision-briefing and prompt-engineering formats are used before they have been hardened, proctoring is recommended (Section 12.3). Batteries should avoid pairing tasks that turn on the same underlying puzzle, such as two scenarios in which the key insight is a measurement artifact, because performance on the second would partly reflect exposure to the first. The interaction transcript and the descriptive interaction metrics are supplementary information and should not be treated as scored dimensions of ability.

14.2 Uses requiring further or local evidence

Three uses require further evidence before they can be recommended. The first is stand-alone use of the short-form tasks in high-stakes decisions, which should await human ratings of candidate responses and the false-accept improvements described in Section 6.4; in the meantime these tasks are suited to inclusion in broader batteries and to developmental use. The second is use of the assessment as an independent hurdle, which, as in Study 3, should be preceded by local evidence linking the tasks to the target role. The third is measurement of change, for example before and after training, which requires parallel forms with demonstrated equivalence; their development is scheduled to begin in October 2026. Until such forms are available, results obtained for developmental purposes should inform discussion rather than be used to estimate change.

14.3 Standard setting and norms

Consistent with Codility's general approach, scores are interpreted against task criteria rather than against a normative population, and no category-wide norms exist. The operational distributions in Section 9.4 can support local benchmarking once non-attempts are separated from attempted responses. Where an employer requires a cut score, it should be set locally, either before administration through a criterion-referenced method such as a modified Angoff procedure with subject-matter experts, or after an initial administration by reviewing score distributions, pass rates, and adverse impact, as described in the *Codility Technical Manual*. Study 3 illustrates why cut scores must be set on the engine's own metric: the

© 2026 Codility. All rights reserved.

correspondence between engine scores and an expert's rating scale is not fixed and must be established for the task and population concerned.

14.4 Limitations of the current evidence

The principal limitations, stated where each result was reported, can be summarized as follows. The evidence of agreement between engine and expert scoring rests on synthetic responses to one scenario that is not in the current library (Study 1) and on authentic responses to one custom task rated by a single expert (Study 3); a two-rater equivalence study on authentic responses has not yet been conducted. Scoring verification identified defects in a substantial proportion of tasks on first pass, and re-test results are not reported here. The short-form tasks have not been evaluated against human ratings of candidate responses, and their synthetic false-accept rates exceed the target. Evidence on generalization across tasks comes from a single task pair, and test–retest and internal-consistency estimates are not yet available. The adverse-impact sample is concentrated in one employer's multinational applicant pool, relies on hypothetical cut scores, and lacks information on country, age, and disability. Criterion-related validity has not been established. The transcript-evaluation evidence concerns engineering transcripts, and its first comparison with expert judgment, based on one expert, did not support the indicator-based methods. The report will be revised as each of these limitations is addressed.

14.5 Research agenda

Table 24 lists the studies planned or in progress, the question each addresses, and its status. Items marked as proposed follow from limitations identified in this report and have not yet been scheduled. Employers deploying the assessment in high-stakes contexts are encouraged to treat their deployments as opportunities for local validation, and Codility's Assessment Science team can support the design of such studies, including content validation against the employer's job analysis, incumbent benchmarking, and criterion studies.

Table 24. Research agenda

Study	Question	Status
Study 3: second hiring round and rescoring	Does the revised scoring key improve agreement with the partner's expert?	In progress
Study 3: criterion-level and criterion-related analyses	Do engine scores agree with the expert's criterion-level ratings, and do they relate to interview outcomes?	Planned; data held by the partner
Two-rater equivalence on authentic responses	Is human–machine QWK at least .70 and within .05 of human–human agreement?	Planned
Holdout replication of Study 1	Does concordance hold for authentic responses not used in calibration, rated by experts independent of development, and does graded credit improve agreement for partially correct responses?	Planned
Short-form task improvement and human ratings	Median false-accept rate below 10%; ICC ≥ .70 on at least 15 of 20 tasks; agreement with human ratings of candidate responses	Scheduled (from October 2026)
Hardening program	Does every live task score below the agreed threshold when an external model answers from the description?	Scheduled (from October 2026)

© 2026 Codility. All rights reserved.

Independent review of medium and hard task tags	Do independent reviewers confirm the criterion-to-skill linkage?	Proposed
Reliability analyses	Internal consistency of criteria; test–retest stability; generalizability across additional task pairs	Proposed
Extended subgroup analyses	Are results comparable across countries, age groups, and operational cut scores?	Proposed
Reading-level review	Is the reading load of business tasks within CEFR B1 bounds?	Proposed
Parallel forms	Do matched forms yield scores within an agreed margin for measurement of change?	Scheduled (from October 2026)
Study 6: accuracy of transcript judgments	Does a structured holistic judgment agree with independent experts as closely as experts agree with each other?	In progress; first single-expert comparison completed

© 2026 Codility. All rights reserved.

Appendix A. Study Summaries

Study 1. Scoring concordance with expert judgment (April 2026)

Element	Summary
Purpose	To determine whether the deterministic engine assigns scores consistent with expert human judgment of written responses.
Design	200 synthetic responses spanning clearly correct, partially correct, and clearly incorrect profiles; ten engine criteria (five lexical, five comparing the response with reference text); three expert raters from the development team scored three dimensions on three-point scales. Follow-up: 50 new synthetic responses scored against a language-model rater.
Principal results	Rater agreement α = .91. Engine vs. consensus: ICC(2,1) = .77, weighted κ = .76, α = .73, r = .80; exact agreement 69%; disagreement beyond one category 1%; F1 .84, .41, and .76 for the low, middle, and high categories; engine mean 0.25 points below raters. Follow-up: initial ICC .49; after correction of defective criteria, ICC .87 (in-sample).
Limitations	Synthetic responses; single scenario not in the current library; raters not independent of development; thresholds calibrated during development; follow-up criterion was a language-model rater.
Next step	Replication on authentic responses not used in calibration, with raters independent of development.

Study 2. Library-wide scoring verification (June 2026)

Element	Summary
Purpose	To verify that each task's scoring behaves as specified in the live candidate environment.
Design	Five tests per task (exemplar, unedited starter, component removal, off-task AI content, external AI answer), applied by eight reviewers to 53 tasks.
Principal results	264 of 265 planned runs completed. Component removal behaved as specified in 52 of 53 tasks; 25 of 53 tasks passed all five tests on first pass; exposure to external AI concentrated in decision-briefing (11 of 15) and prompt-engineering (7 of 9) tasks.
Limitations	Engineering and content verification rather than psychometric evidence; no fixed numeric threshold for the external-AI test; re-test results not reported.
Next step	External-AI testing of every live task under the hardening program.

Study 3. Field study in live hiring (June to September 2026)

Element	Summary
Purpose	To compare engine scores with expert ratings of authentic applicant responses to a custom task in live hiring.
Design	Large U.S. telecommunications company; senior analyst role in customer relations; task built from the partner's job analysis; 75 minutes; twelve criteria; one trained partner expert rated every response.

© 2026 Codility. All rights reserved.

Element	Summary
Principal results	96 applicants; 29 hires. Mean engine score by expert rating: 37%, 60%, 76%, 83%, 87%; ρ = .72; τb = .63; pass/fail agreement 79% at exploratory cuts; QWK on uncalibrated bands .50.
Limitations	One rater and no human–human baseline; one task; bands defined after the fact; expert ratings may not be independent of hiring decisions.
Next step	Rescoring with the revised key; criterion-level and interview-outcome analyses; two-rater equivalence study.

Study 4. Expert review of the short-form tasks (July to September 2026)

Element	Summary
Purpose	To screen the twenty short role-agnostic tasks for construct fit, defensibility of the scoring key, fairness concerns, and function in the platform before wider validation.
Design	Synthetic concordance against a language-model judge (validation only); review by three assessment scientists under decision rules fixed in advance, with independent skill naming and in-platform attempts; two reviewers provided structured ratings.
Principal results	7 of 10 lean tasks met all three concordance targets; 20 of 22 construct-fit and 19 of 22 defensibility ratings at 4 or 5 of 5; all 20 tasks cleared, 8 after targeted scoring changes; false-reject rates 0% to 9.1%.
Limitations	No candidate sample; four tasks without human construct ratings; median synthetic false-accept rate about 25%; 4 of 20 tasks at ICC ≥ .70 against the AI rater.
Next step	Reduction of false accepts below 10%; human ratings of candidate responses.

Study 5. Subgroup differences and candidate reactions (July to October 2026)

Element	Summary
Purpose	To estimate subgroup differences in scores and selection rates and to describe candidates' reactions.
Design	Impact ratios against the four-fifths guideline, Fisher and Z tests, Mantel–Haenszel tests, one-person flip rule, Hedges' g, Welch and fixed-effects models; Holm-corrected comparisons of perceptions; independent replication in Python.
Principal results	1,006 sessions, 330 with demographics. Gender differences negligible; no adverse impact against women; |g| ≤ 0.37 across adequately sized comparisons; impact ratios below .80 for White and Hispanic or Latino candidates relative to Asian candidates at cut scores of 60 and 70; no gender difference in perceived fairness (composite α = .85).
Limitations	Hypothetical cut scores; one employer's multinational pool predominant; respondents scored slightly higher than non-respondents; unscored tasks counted as zero in the combined score (48 of 1,006 sessions; 6 of 330 with demographics); no country, age, or disability data.
Next step	Country and age analyses; analysis at operational cut scores; ongoing monitoring.

© 2026 Codility. All rights reserved.

Study 6. Transcript-evaluation methods (July to October 2026, interim)

Element	Summary
Purpose	To evaluate methods for judging AI interaction transcripts, including their agreement with expert judgment.
Design	Context engineering judged by holistic and indicator-based language-model methods in repeated runs: 200 and 300 sessions from an external corpus of AI-assisted engineering work; 92 and 233 Codility engineering candidates; holistic ratings by one expert for 61 candidates. Reproducibility of displayed evidence for five ESM AI skills.
Principal results	Run-to-run ICC(3,1) .85 to .95; classification changes 30.8% (holistic) vs. 21.0% (indicator-based) in the first phase; no relation to session success; ceiling on indicator banks; scores related to amount written (r up to .72). Agreement with the expert: ICC(A,1) .63 for the holistic rating and .11 to .26 for indicator-based methods; pre-set criteria not met by the indicator-based level method. Displayed evidence: κ .83 to .88 across runs.
Limitations	Engineering transcripts rather than AI for Business sessions; one construct; one expert and one task in the expert comparison; AI-generated success criterion; evidence correctness checked only in internal audits.
Next step	Structured holistic design validated against fresh ratings from two experts; blind review of displayed evidence; extension to AI for Business transcripts.

© 2026 Codility. All rights reserved.

Appendix B. Task Evidence Register

Table B1 lists each task in the standard AI for Business library as of October 2026. For each task it gives the format, the number of scored criteria, the verification study in which the task was examined, and the number of scored administrations in the July to October 2026 operational data used in Section 9 and Study 5. Detailed verification results for individual tasks are available on request.

Table B1. Task evidence register (71 tasks)

#	Task	Role family	Format	Criteria	Verification	Operational n
1	Activation Decay	Customer success	Research drilldown	5	Study 2	—
2	Billing Dispute	Customer success	AI output audit	9	Study 2	—
3	Quiet Slide	Customer success	Research drilldown	5	Study 2	—
4	Response Brief	Customer success	Prompt engineering	20	Study 2	—
5	Risk Queue	Customer success	AI output audit	12	Study 2	—
6	AI Assist Paradox	Customer support	Decision briefing	9	Study 2	—
7	Closed Loop	Customer support	Research drilldown	5	Study 2	—
8	Outage Dip	Customer support	AI output audit	12	Study 2	—
9	Phantom Bug	Customer support	Decision briefing	9	Study 2	—
10	Root Cause Audit	Customer support	AI output audit	12	Study 2	—
11	SLA Trail	Customer support	Research drilldown	5	Study 2	—
12	Adoption Audit	Data analysis	AI output audit	12	Study 2	—
13	Blended Rate	Data analysis	Prompt engineering	20	Study 2	—
14	Cohort Dilution	Data analysis	Decision briefing	12	Study 2	—
15	Forecast Fracture	Data analysis	Research drilldown	5	Study 2	—
16	Phantom Lift	Data analysis	Decision briefing	12	Study 2	76
17	Schema Shift	Data analysis	Research drilldown	5	Study 2	—
18	Clean Close	Finance	AI output audit	12	Study 2	—
19	Close Variance Prompt	Finance	Prompt engineering	20	Study 2	—
20	H1 Revenue Miss	Finance	Research drilldown	5	Study 2	—
21	Opex Overrun	Finance	Decision briefing	12	Study 2	—
22	Q4 Forecast Pick	Finance	Decision briefing	12	Study 2	—
23	Reforecast Brief	Finance	Prompt engineering	20	Study 2	—
24	AI Adoption	General workforce	Decision briefing	12	Study 2	770
25	Budget Audit	General workforce	AI output audit	12	Study 2	716
26	Compliance Prompt	General workforce	Prompt engineering	20	Study 2	—
27	Meeting Tool	General workforce	Decision briefing	12	Study 2	—
28	Office Relocation	General workforce	Research drilldown	5	Study 2	136
29	Supplier Dispute	General workforce	Prompt engineering	20	Study 2	23
30	Attribution Blind Spot	Marketing	Prompt engineering	20	Study 2	—

© 2026 Codility. All rights reserved.

31	Brand Brief	Marketing	Research drilldown	5	Study 2	—
32	Channel Mix	Marketing	AI output audit	12	Study 2	—
33	Growth Burst	Marketing	Decision briefing	12	Study 2	—
34	Retargeting Saturation	Marketing	Decision briefing	12	Study 2	—
35	Segment Audit	Marketing	AI output audit	12	Study 2	—
36	AI Feature Retention	Product management	Decision briefing	9	Study 2	—
37	Data Drilldown	Product management	Research drilldown	5	Study 2	—
38	Feasibility Triage	Product management	Research drilldown	5	Study 2	—
39	Market Share Myth	Product management	AI output audit	12	Study 2	—
40	Roadmap Bet	Product management	Decision briefing	12	Study 2	—
41	AI Positioning Bet	Product marketing	Decision briefing	9	Study 2	—
42	Competitive Prompt	Product marketing	Prompt engineering	20	Study 2	—
43	NPS Paradox	Product marketing	Decision briefing	9	Study 2	—
44	Positioning Intel	Product marketing	Research drilldown	5	Study 2	—
45	Positioning Report	Product marketing	AI output audit	12	Study 2	—
46	Dark Deal	Sales	Research drilldown	5	Study 2	—
47	Deal Priority	Sales	Decision briefing	9	Study 2	—
48	Fast Brief	Sales	Prompt engineering	20	Study 2	—
49	Integration Alibi	Sales	Research drilldown	5	Study 2	—
50	Phantom Pipeline	Sales	AI output audit	9	Study 2	—
51	Win/Loss Report	Sales	AI output audit	12	Study 2	—
52	The Agenda	General workforce	Lean: write a prompt	5	Study 4	—
53	The Decline	General workforce	Lean: write a prompt	5	Study 4	—
54	The Status Line	General workforce	Lean: write a prompt	5	Study 4	—
55	The One-Screen Summary	General workforce	Lean: write a prompt	5	Study 4	—
56	Notes to Actions	General workforce	Lean: write a prompt	5	Study 4	—
57	The Blunt Email	General workforce	Lean: correct a draft	5	Study 4	—
58	The Long Way Round	General workforce	Lean: correct a draft	5	Study 4	—
59	The Wrong Date	General workforce	Lean: correct a draft	5	Study 4	—
60	The Half Answer	General workforce	Lean: correct a draft	5	Study 4	—
61	The Jargon Wall	General workforce	Lean: correct a draft	5	Study 4	—
62	The Monday Morning Pile-Up	General workforce	Inbox reply	8	Study 4	—
63	The Double-Booked Executive	General workforce	Inbox reply	8	Study 4	—
64	The Rumor Mill	General workforce	Inbox reply	8	Study 4	—
65	The Vendor Squeeze	General workforce	Inbox reply	8	Study 4	—
66	The Credit Grab	General workforce	Inbox reply	8	Study 4	—
67	The Cascading Deadline	General workforce	Inbox reply	8	Study 4	—
68	The Ambiguous Ask	General workforce	Inbox reply	8	Study 4	—
69	The Overflowing Calendar	General workforce	Inbox reply	8	Study 4	—
70	The Data Discrepancy	General workforce	Inbox reply	8	Study 4	—

© 2026 Codility. All rights reserved.

71	The Well-Meaning Firehose	General workforce	Inbox reply	8	Study 4	—

© 2026 Codility. All rights reserved.

Appendix C. Methods Notes

C.1 Agreement statistics

ICC(2,1) is the two-way random-effects, absolute-agreement, single-rater intraclass correlation, interpreted according to Koo and Li (2016). Quadratically weighted κ (QWK) weights disagreements by the square of their distance and is the conventional agreement statistic in research on automated scoring (Williamson et al., 2012). Ordinal Krippendorff's α is interpreted against the .667 minimum suggested for tentative conclusions (Krippendorff, 2004). Spearman's ρ and Kendall's τb index rank-order association; τb corrects for ties.

C.2 Study 3 cross-tabulation and statistics

Table C1 reproduces the cross-tabulation of the partner expert's overall rating against engine scores grouped into five equal-width bands, from which the statistics in Section 8.3 and Table C2 were computed.

Table C1. Study 3: Partner expert rating by engine score band (N = 96)

Expert rating	Band 1 (0–20%)	Band 2 (21–40%)	Band 3 (41–60%)	Band 4 (61–80%)	Band 5 (81–100%)	Total
1	2	5	5	0	0	12
2	0	2	12	12	2	28
3	0	1	3	7	13	24
4	0	0	2	3	14	19
5	0	0	0	2	11	13
Total	**2**	**8**	**22**	**24**	**40**	**96**

Table C2. Study 3: Association and agreement statistics computed from the banded data

Statistic	Value	Note
Spearman ρ	.72	Rank-order association
Kendall τb	.63	Rank-order association, corrected for ties
Pass/fail agreement (expert ≥ 3) at engine ≥ 61% / ≥ 81%	79% / 79%	κ = .56 / .59; cut points examined after the fact
Mean band, engine vs. expert	3.96 vs. 2.93	Bands not calibrated to the expert's scale
Exact band agreement	22%	Reduced by the offset between bands and scale
Agreement within one band	67%	
Quadratic weighted kappa	.50	Uncalibrated equal-width bands; not an equivalence test

© 2026 Codility. All rights reserved.

C.3 Adverse impact methods

Impact ratios compare the lower selection rate with the higher and are evaluated against the four-fifths guideline (*Uniform Guidelines*, 1978). Statistical significance was assessed with pooled two-proportion Z tests and two-tailed Fisher's exact tests (Morris & Lobsenz, 2000). The one-person flip rule asks whether an impact ratio below .80 would reach .80 if one additional member of the lower-rate group passed. Mantel–Haenszel tests stratified by test version guard against aggregation artifacts across batteries (Mantel & Haenszel, 1959). Comparisons of perceptions used Mann–Whitney U tests with Holm correction across eleven variables (Holm, 1979). Minimum detectable effects assume α = .05 and power of .80.

C.4 Spearman-Brown projections

The projected consistency of a composite of k tasks is k·r / (1 + (k − 1)·r), where r is the observed correlation between tasks (r = .405, n = 716). The projection assumes that additional tasks are parallel to those observed; because the two tasks differ in format and content, the projections are approximate.

© 2026 Codility. All rights reserved.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. AERA.

Anthropic. (2026). *2026 agentic coding trends report*. Anthropic.

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Campion, M. A. (2026). Can legal and professional personnel selection principles be met with machine learning (artificial intelligence)? *Human Resource Management, 65*, 235–255.

Campion, M. A., & Campion, E. D. (2026). *How to evaluate artificial intelligence (AI) based employment tools from vendors*. SIOP Foundation and CHRO Association.

Campion, M. A., Fink, A. A., Ruggeberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B. (2011). Doing competencies well: Best practices in competency modeling. *Personnel Psychology, 64*(1), 225–262.

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101*(7), 958–975.

Codility. (2026a). *Codility technical manual* (Version 3.1). Codility.

Codility. (2026b). *Codility Engineering Skills Model technical report* (Version 2.1). Codility.

Codility. (2026c). *NLP scoring validation report: Active Accounts assessment* (Version 1.0) [Internal report]. Codility.

Codility. (2026d). *Synthetic NLP validation of the Active Accounts task: Results and next steps* [Internal report]. Codility.

Codility. (2026e). *Context engineering scoring prototype: Technical methods and interim validation results* [Internal draft technical manual]. Codility.

Codility. (2026f). *Candidate AI transcript scoring: 233-candidate baseline for expert review* [Internal report]. Codility.

Codility. (2026g). *Expert evaluation study plan: AI transcript scoring* (Draft 1) [Internal document]. Codility.

Codility. (2026h). *N300 SWE-chat context engineering study: Results and findings* [Internal report]. Codility.

Codility. (2026i). *Real Codility candidate scoring study: Results and findings* [Internal report]. Codility.

Codility. (2026j). *Context engineering level versus expert ratings: Results* [Internal report]. Codility.

Codility. (2026k). *AI transcript evidence: Reliability evidence* [Internal report]. Codility.

Dakan, R., & Feller, J. (2025). *AI fluency: Frameworks and foundations*. Anthropic.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*(4), 694–734.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*(3), 639–683.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

© 2026 Codility. All rights reserved.

Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., … Newton, C. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. *Personnel Psychology, 76*(4), 1061–1123.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Liff, J., Mondragon, N., Gardner, C., Hartwell, C. J., & Bradshaw, A. (2024). Psychometric properties of automated video interview competency assessments. *Journal of Applied Psychology, 109*(6), 921–948.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Morris, S. B., & Lobsenz, R. E. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology, 53*(1), 89–111.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, 100041.

Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58*(4), 1009–1037.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology, 107*(11), 2040–2068.

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.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215–232.

Uniform guidelines on employee selection procedures. (1978). 29 C.F.R. Part 1607.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13.

© 2026 Codility. All rights reserved.

Version History

Version	Date	Description
1.0	July 2026	Initial release of the AI for Business Technical Report
2.0	October 2026	Reorganized as a validity argument. Added library-wide scoring verification (Study 2), a field study in live hiring (Study 3), expert review of the short-form tasks (Study 4), analyses of subgroup differences and candidate reactions (Study 5), and the research on transcript evaluation (Study 6). Revised the account of scoring concordance (Study 1) from the primary validation report and added its follow-up analysis. Added sections on content-related evidence, reliability and generalizability, language and test security, and intended use and limitations, together with a task evidence register and methods notes. Updated validity estimates and references.

© 2026 Codility. All rights reserved.