2

How Are Subscores Reported?

> There is no such thing as information overload. There is only bad design.
>
> *Edward Tufte*

Daisy, a 10-year-old fifth grader at Lakeville Elementary school, is a bright, enthusiastic, hard-working student. Her parents have set an extremely high standard for her education and thus, although Daisy consistently performed well above average among her classmates, her family was concerned that she was not reaching her full potential. Without any prompts from the school, pediatrician, or other care providers, Daisy's parents brought her to a psychologist to be assessed for a learning disability, specifically asking to check if she was behind in her reading and vocabulary. As requested, the psychologist administered a battery of relevant assessments and provided the parents with a report summarizing the results. Contrary to her parents' concern, Daisy scored very well in reading and vocabulary. However, the subscores on one of the assessments happened to be normed within-person, designed to identify an individual's relative strengths and weaknesses by contrasting their subscore performance against their overall assessment performance. In school Daisy always performed above average in math, but on this particular assessment, due to her extremely high reading proficiency, it now appeared as if Daisy had a weakness in math.

Her parents were ecstatic; this was exactly what they had hoped for, as they interpreted Daisy's relative weakness as evidence that she had a learning disability and needed additional support. Daisy's parents built a case around this single math subscore to try to pressure the school into qualifying her for an individualized education plan, which would grant curriculum

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

and testing accommodations. Given Daisy's stellar academic performance, the school resisted, as they would rather reserve these essential resources and services for struggling students who were truly in need. However, Daisy's parents were very demanding – threatening litigation – and the school district was mindful of the broader consequences within the community if the issue was not quickly resolved. Now with mounting pressure from the district, the school reluctantly granted Daisy's parents the requested accommodations, thus formalizing a learning disability in her academic record. With her newly diagnosed learning disability, Daisy was protected under the Individuals with Disabilities Education Act (IDEA), and her school, as well as future schools that she would attend, was required by federal law to provide a variety of entitlements and interventions.

Daisy's is a fictionalized account of a true story and reflects a situation that has been worsening over time as more and more parents and students interested in an academic advantage seek diagnoses to qualify for educational and testing accommodations (Lovett & Harrison, 2021). This story highlights the intricacy of subscore reporting by illustrating how subscores can be misinterpreted and misused, leading to serious practical consequences. Reporting a subscore is no different from reporting a total score, and thus the same principles for best practices in score reporting are applicable. In their comprehensive review of student score reports, Goodman and Hambleton (2004) noted several reoccurring challenges around how results are presented, such as too much statistical jargon, lack of descriptive information, and dense tables and graphs. Reporting subscores adds a layer of complexity to these existing concerns because they often contain poor psychometric properties.

In Chapter 1 we described what a subscore is, why there is a desire to report them, and a sense of the effort involved when they need to be accurate. But exactly how subscores are reported varies based on several important considerations, including the psychometric quality of the data underlying the subscores, the inferences that need empirical support, the type of assessment, and the nature of the users who will be receiving and acting upon the subscore information.

We will cover how to determine when subscores are reportable from a statistical perspective in Chapter 3. Additionally, the conditions that impact the psychometric value of a subscore and what to do when they are not worth reporting will be covered in Chapters 4 and 5. Thus, for this chapter, we will assume that the reported subscores have at least some value. Depending on that value, as well as other considerations discussed later in this chapter, there are many ways to present the same subscore results.

2.1 Subscore Value

For a hypothetical student and test composed of five content areas, Figure 2.1 illustrates how the same subscore information can be presented in a few different ways: (a) raw scores, (b) percent correct and percentile scores, (c) profile bands, or (d) categorical performance indicators.

Figure 2.1a is the most detailed at the subtest level and provides raw information that may be the simplest for a score user to interpret, making transparent how many items are in each content area and the number of correct responses. In Figure 2.1b, percent correct scores instead of raw scores are reported, which obscures the number of items and makes it difficult for the score user to know how many items were responded to incorrectly. However, this can be advantageous in facilitating comparisons among categories if there are substantive

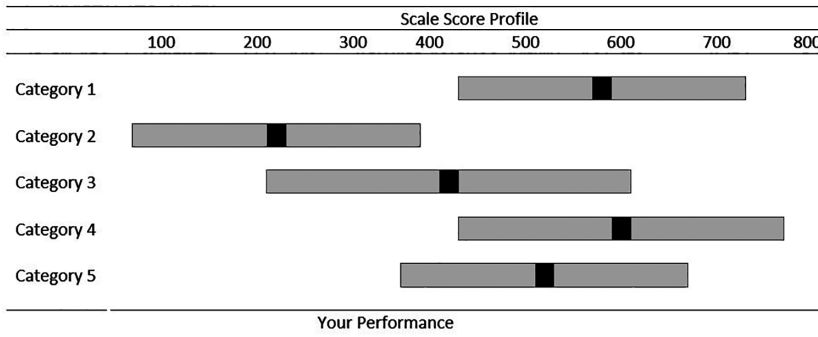
(a)

Content Area	Number Correct
Category 1	31 out of 44
Category 2	15 out of 43
Category 3	14 out of 26
Category 4	28 out of 39
Category 5	28 out of 43

(b)

Content Area	Percent Correct Score	Percentile Score
Category 1	70	89
Category 2	35	48
Category 3	54	61
Category 4	72	92
Category 5	65	74

(c)



(d)

Content Area	Low Performance	Average	High Performance
Category 1			X
Category 2	X		
Category 3		X	
Category 4			X
Category 5		X	

Figure 2.1 A few illustrations of the variety of ways to report subscores. (a) Raw scores; (b) percent correct and percentile scores; (c) profile band; (d) categorical performance.

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

discrepancies in the number of items, sometimes referred to as weighting, between content areas. For instance, a raw score of 9 points should be interpreted differently if it was out of 10 total points compared to 30 total points. Additionally, the normative information in Figure 2.1b adds context for how users can understand their performance as a percentile relative to some meaningful comparison group. For instance, 89% of the students in the school district scored lower than the student in Category 1.

Instead of numeric scores, Figure 2.1c provides profile bands – sometimes referred to as performance bands, confidence bands, or score intervals – with the placement of the students' numeric scale scores marked in the center. Traditionally, the width of the interval is an illustration of measurement error, with narrower intervals representing greater accuracy and wider intervals having lesser accuracy. Presenting subscore information in this format can provide a more direct sense of imprecision to help mitigate overinterpretation of minor score differences between categories – usually accompanied by text explaining that overlapping bands should be interpreted as similar performance.

Lastly, Figure 2.1d altogether removes the burden of asking users to interpret measurement errors and instead reports discretized categories that incorporate the measurement error to identify whether an examinee's performance differed from a relevant reference point. Although useful to facilitate interpretations, it should be noted that the categorizations also contain error, can sometimes be less reliable than the scores from which they are constructed (Ramsay, 1973), and can even be used fraudulently to skew the results (Wainer, Gessaroli, & Verdi, 2006). Further, the consequences of either random or systemic error that puts an examinee's true score in the wrong category might be more of a risk compared to a more subtle difference of a few points. Thus, caution is needed when defining thresholds for determining how continuous scores are converted into discretized categories, which is no different from specifying pass/fail on a classification test (Angoff, 1971) or implementing indicators of subscore performance (Feinberg & von Davier, 2020) as low, average, or high as illustrated in Figure 2.1d.

Figure 2.1a–d also reflect a continuum of decreasing granularity of the subscore information. There are, of course, other modalities of presenting subscore results not displayed here, some of which will be illustrated later in this chapter when describing sample reports from operational testing programs. Generally, more specificity can be supported for subscores of higher precision and accuracy, which in turn helps to mitigate misinterpretation. However, there could be other scoring implications that make reporting the more intuitive raw or percent correct scores less appropriate than a scale score that has gone through further processing. For instance, a student may have a lower raw score

in Category A compared to Category B, but if the items in Category A were more difficult, then a lower raw or percent correct score may not necessarily reflect a lower underlying proficiency. In such cases, an adjustment is needed to correct for difficulty, which could be done between categories on the test as well as across different forms of the test (alternatively, report users should be explicitly cautioned about differences in difficulty). Another example is when the reported score has been scaled relative to a particular group of interest, often referred to as a norming group, which could be any subgroup relevant to the score users such as a school, state, national population, first-year college students, successful job applicants, or individuals diagnosed with posttraumatic stress disorder (PTSD) (Mertler, 2018). A normed scale score provides context and supports inferences against the relevant comparison group. Thus, a seeming loss of granularity can be offset by providing more subscore utility.

2.2 Types of Assessment

Beyond psychometric properties and needs for other statistical and scaling adjustments, considerations for how to display subscore information also depend on the type of assessment and corresponding inferences the test publisher claims users can make with the subscores. For our purposes, we will refer to two broad categories of assessments, formative and summative

Formative assessments are often low-stakes and used within educational programs to identify what students misunderstand in an effort to improve instruction and learning (Perie, Marion, & Gong, 2009). At an individual test-taker level, formative assessment is designed to provide feedback to the user, typically on their relative strengths and weaknesses against a relevant criterion. In terms of purpose, formative assessment total scores and subscores serve to inform remediation for learning and preparation for future assessments. The criterion for this feedback could be the user (e.g., interpret performance relative to oneself), a norm group (e.g., interpret performance relative to other students in a similar training program, year of matriculation, proficiency level), or a standard (e.g., interpret performance relative to grade level expectations set by the state). On an aggregate level, the criterion could be how the students at one school performed relative to other cohorts within or across institutions, though the focus would principally be on school-level remediation, such as improving curriculum. There are different types of formative assessments, such as a practice test, a self-assessment, or a progress test that an individual may repeat periodically to track their improvement over time. These types include formative assessments that users may intentionally not prepare for or complete at the

start of their learning journey if they are interested in their baseline knowledge or performance. As the primary focus of formative assessment is on supporting improvement, careful consideration should be given to whether and how results are shared with other decision-makers to reduce concerns of having academic consequences for poor performance.

Summative assessments are high-stakes and often shared with decision-makers because they are designed for evaluation (Perie, Marion, & Gong, 2009). Examples of summative assessments include admission tests used in a selection process, achievement tests to assess learning of a particular construct, licensure and certification tests designed to classify test takers as either passing or failing, or national assessments linked to defined accountability standards. Though the main results of a summative assessment are intended to be communicated to stakeholders beyond the examinee to support consequential decisions, subscores are commonly included to support remediation and help users better understand their performance. Thus, though subscores naturally align within a formative assessment context, they can be a major component of how results are reported on summative assessments. For assessing broader group-level trends, such as on the National Assessment of Educational Progress (NAEP), subscores are aggregated over different subsets of students to support various targeted inferences such as by type of school, ethnicity, gender, geographical location, race, and disability status (NAEP, 2023).

However, because summative assessments are designed for evaluation, building quality subscores to support formative uses is often deprioritized in the tradeoff to maximize the primary inference. When this does occur, test publishers may attempt to find a compromise in reporting coarser subscore information (e.g., performance intervals or categories rather than numeric scores) to reduce misinterpretation. Thus, many of the concerns of reporting subscores with poor psychometric value stem from summative assessment programs that are in a difficult situation of trying to both satisfy the primary inference for the overall test score and support score users who desire diagnostic feedback.

2.3 Types of Users

When we think of a standardized assessment, we often first envision high school students sitting for the SAT after months of preparation and managing the stress that often accompanies knowing their performance qualifies them to get into their desired college. These users, though hopefully knowledgeable about the quantitative content on the test, would likely (1) not have any expertise in interpreting complex subscores and (2) be highly motivated to receive detailed

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

diagnostic feedback to understand their performance. In this instance, students eagerly awaiting their SAT results are no different from premed students after completing the Medical College Admissions Test (MCAT) or aspiring accountants completing the Certified Public Accounting (CPA) examination. To assist with score interpretation, test publishers commonly include explanatory language on the score report as well as other supplementary materials (e.g., a video walkthrough of interpreting a score report on their website). Additionally, these users may share their score reports with teachers, faculty, advisors, or learning specialists to help them interpret and make their results actionable. Test publishers would also want to design and refine the report by engaging a representative sample of stakeholders with surveys, cognitive interviews, and focus groups so as to ensure that users can correctly interpret the results.

In an aggregate case, when subscores may be compiled across students within a program or institution, the user may be a teacher, a program director, or another institutional administrator looking for feedback to make curricula enhancements. This could also include governing boards or agencies if results are compiled across schools and districts to inform higher-level decisions about policies, funding, or educational reform. These types of users, faculty, administrators, and government representatives routinely interpret score reports and thus may require less detailed score interpretation materials or protection from misinterpretation.

Similarly, psychologists who frequently administer assessments often score them manually and summarize the results, consulting a manual only for general guidance on the scores when needed. In some cases, the test publisher may provide only a scoring rubric, as the psychologist would be compiling a comprehensive summary, usually across multiple assessments, for a client that also includes background information, behavior observations, and their clinical interpretation.

Thus, in addition to the varying psychometric quality of subscores, many types of assessments report subscores for different purposes and communicate them to different types of users, all of which are factored into how a testing publisher determines what to report. The next section will review a few sample score reports to illustrate how this appears in practice.

2.4 Score Report Examples

2.4.1 Praxis®

The Praxis is a certification test used by many states as one of several requirements to become a certified teacher in the United States (PRAXIS, 2023). Figure 2.2 presents an excerpt from a sample score report from Sinharay et al.

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

Test / Test Category *	Your Raw Points Earned	Average Performance Range **
ELEMENTARY EDUCATION: CURRICULUM, INSTRUCTION, AND ASSESSMENT (5017)		
I. READING AND LANGUAGE ARTS	33 out of 37	23–29
II. MATHEMATICS	26 out of 31	19–25
III. SCIENCE	15 out of 20	11–15
IV. SOCIAL STUDIES	14 out of 17	9–13
V. ART, MUSIC, AND PHYSICAL EDUCATION	13 out of 15	6–12

* Category-level information indicates the number of test questions answered correctly for relatively small subsets of the questions. Because they are based on small numbers of questions, category scores are less reliable than the official scaled scores, which are based on the full sets of questions. Furthermore, the questions in a category may vary in difficulty from one test to another. Therefore, the category scores of individuals who have taken different forms of the test are not necessarily comparable. For these reasons, category scores should not be considered a precise reflection of a candidate's level of knowledge in that category, and ETS recommends that category information not be used to inform any decisions affecting candidates without careful consideration of such inherent lack of precision.

** The range of scores earned by the middle 50% of a group of test takers who took this form of the test at the most recent national administration or other comparable time period. N/C means that this range was not computed because fewer than 30 test takers took this form of the test or because there were fewer than eight questions in the category or, for a constructed-response module, fewer than eight points to be awarded by the raters. N/A indicates that this test section was not taken and, therefore, the information is not applicable.

Figure 2.2 Excerpt from a Praxis score report. Copyright © 2023 by Educational Testing Service (ETS). Reproduced with permission. All rights reserved.

(2019) showing how the Praxis subscore information is reported along with the corresponding explanatory text. Subscore information is conveyed at a very granular level, similar to Figure 2.1a, along with normative information on how the middle 50% of a recent group of test takers from a national administration performed on this form of the test. Based on these results, the recipient can see that they performed above average in all categories. Going further, the recipient might infer that mathematics and science are areas to focus on for additional preparation, given that they missed five points in each category. However, the explanatory text in the footnote discourages making any serious inferences due to the quality of the subscores: "For these reasons, category score should not be considered a precise reflection of a candidate's level of knowledge in that category, and ETS recommends that category information not be used to inform any decisions affecting candidates without careful consideration of such inherent lack of precision." Interestingly, the footnote also references symbols of N/C for not computed and N/A for not applicable, but neither are observed in the report. Most likely this is standard language on a template used for multiple forms of the Praxis, though it could still be confusing to score users who wonder if something is missing or even why the reference is included at all.

2.4.2 SAT

The SAT is a summative assessment designed to inform college admission decisions and reports subscores on the math and evidence-based reading and writing sections. Similar to how Praxis scores by themselves do not qualify someone to be a teacher, there is no minimum acceptable SAT score, as


CollegeBoard SAT

SAT Score Report

Test Date: **March 13, 2021**
Registration Number: **0123456789**
Sex: **Female**
Date of Birth: **Feb. 12, 2004**
Test Center Number: **12345**
CB Student ID: **12345678**
High School Code: **123456**
High School Name: **John F. Kennedy High School**

Imagood Student
100 Main Street
Apt 2
Anytown, MA 00000-0000

Your Total Score
1010 | *400–1600*

50th Nationally Representative Sample Percentile
41st SAT User Percentile

Section Scores

490 | *200–800*
Your Evidence-Based Reading and Writing Score
44th Nationally Representative Sample Percentile
35th SAT User Percentile
You've met the benchmark!

520 | *200–800*
Your Math Score
57th Nationally Representative Sample Percentile
49th SAT User Percentile
You've scored below the benchmark.

Test Scores

22 | *10–40* Reading
27 | *10–40* Writing and Language
26.0 | *10–40* Math

Cross-Test Scores | *10–40*
24 Analysis in History/Social Studies
23 Analysis in Science

Subscores | *1–15*
6 Command of Evidence
9 Words in Context
8 Expression of Ideas
9 Standard English Conventions
8 Heart of Algebra
9 Problem Solving and Data Analysis
7 Passport to Advanced Math

Am I on Track for College?
Look for the green, yellow, or red symbols next to your section scores. They let you know if your scores are at or above the benchmark scores. Benchmarks show college readiness. If you see green, you're on track to be ready for college when you graduate.
If you score below the benchmark, you can use the feedback and tips in your online report to get back on track.
Benchmark scores:
Evidence-Based Reading and Writing: 480
Math: 530

How Do My Scores Compare?
A percentile shows how you scored, compared to other students. It's a number between 1 and 99 and represents the percentage of students whose scores are equal to or below yours.
For example, if your Math percentile is 57, that means 57% of test takers have Math scores equal to or below yours.
The Nationally Representative Sample Percentile compares your score to the scores of typical U.S. students.
SAT® User Percentile compares your score to the scores of students who typically take the test.

How Can I Improve?
To see which skills are your strongest and what you can do to boost your college readiness, go to your full report online and look for Skills Insight™.

What Are Score Ranges?
Test scores are single snapshots in time—if you took the SAT once a week for a month, your scores would vary.
That's why score ranges are better representations of your true ability. They show how much your score can change with repeated testing, even if your skill level remains the same.
Colleges know this, and they get score ranges along with scores so they can consider scores in context.
Your online score report shows your score ranges.

Figure 2.3 Excerpt from an SAT score report. © Copyright 2021 College Board. "SAT Suite Results: 2020 (State Reports)." All rights reserved. Reprinted with permission.

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

different colleges and universities apply different thresholds and use the SAT as a component within a comprehensive admissions process. Figure 2.3 presents an excerpt from a sample SAT score report that includes an individual's total score, performance on different sections, and then subscores at the bottom (SAT, 2023). Subscores on the SAT report are presented as scale scores but with little additional score information. Instead, language on the report directs students online, where they can find a more interactive experience that allows them to explore their skills and connect to appropriate remediation resources if desired. Thus, based on only this report, a recipient can compare their performance between categories, but they would lack context relative to a meaningful group or the significance of the observed scaled score differences (e.g., whether a two-point difference is worthy of attention).

2.4.3 ACT

Like the SAT, the ACT is also a summative assessment that is primarily designed to inform college admission decisions, but it defines the construct slightly differently and reports subscores from the English, math, reading, science, and writing sections of the test. Figure 2.4 presents an excerpt from a sample ACT score report (ACT, 2023). Subscore information includes the number correct out of the total number of items in a particular content area, the corresponding percentage correct, and then a visual indicator representing whether the student's performance was in the readiness range – which is also communicated with a checkmark. As indicated in the explanatory text at the bottom of the report, the readiness range "shows where a student who has met the ACT College Readiness Benchmark on this subject test would typically perform." The explanatory text also notes that meeting the readiness range indicates that "you have at least a 50% chance of obtaining a B or higher or about a 75% chance of obtaining a C or higher in specific first-year college courses in the corresponding subject area." A recipient of this report can see that all their science and all but one of their math subscores were below the readiness range and thus are in need of remediation if they view these grade probabilities as a meaningful goal. This interpretation also aligns with their US and state rankings that illustrate mostly below-average performance in math and science.

2.4.4 American Board of Internal Medicine Maintenance of Certification (ABIM MOC)

The American Board of Internal Medicine Maintenance of Certification Examination (ABIM MOC) is a classification test for practicing physicians to

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

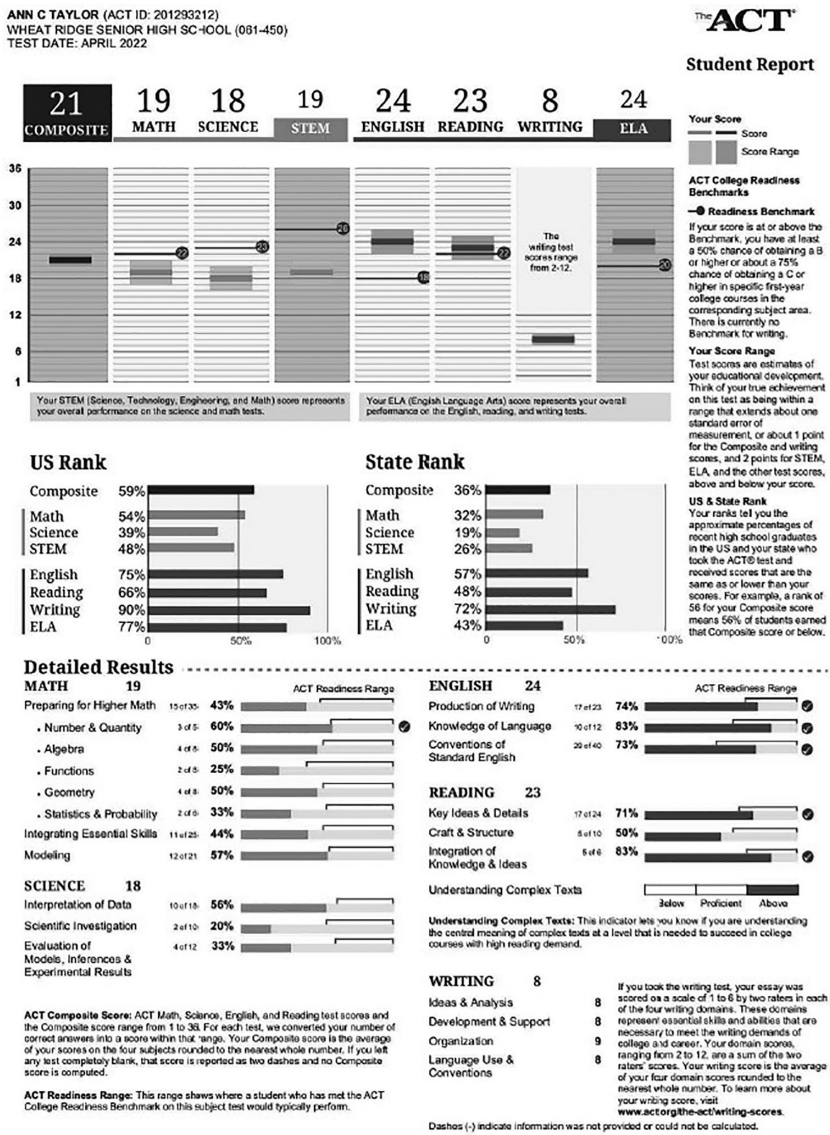


Figure 2.4 Excerpt from an ACT score report. Copyright © 2021 by ACT. Reproduced with permission. All rights reserved.

maintain their specialty credentials by completing and passing the test every 10 years. Figure 2.5 presents an excerpt from the second page of a sample ABIM MOC score report (ABIM MOC, 2023). Subscores for each medical

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

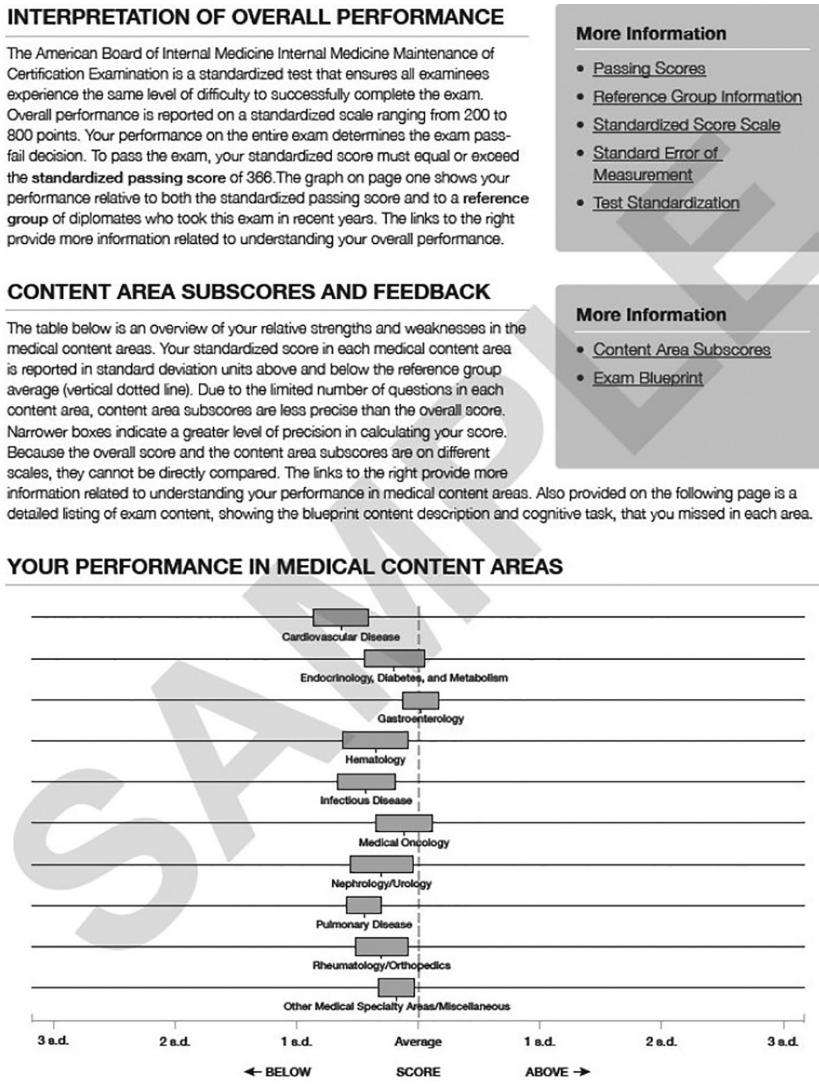


Figure 2.5 Excerpt from the second page of the ABIM MOC score report. Reproduced with permission. All rights reserved.

content area are reported as profile bands standardized to a recent group of diplomates, with a vertical line indicating average performance. As noted in the interpretive text above the graph and similar to Figure 2.1c, the width of the bands is a reflection of score precision. Subsequent pages of the report, not shown here, provide item-level information for missed content as a brief

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

phrase descriptive of the concept targeted by the item. The recipient of this report would see that they performed slightly below average across most subscores. Given the overlap across profile bands, there are no specific areas for remediation; rather, the recipient would benefit from additional preparation on every aspect of the content domain.

2.4.5 Armed Services Vocational Aptitude Battery (ASVAB)

As we discussed in Chapter 1, the Armed Services Vocational Aptitude Battery (ASVAB) is a summative assessment designed to inform military entrance eligibility and help determine what branch of service would be a good fit for an applicant. Figure 2.6 displays an excerpt from a sample score report that individuals receive after completing the ASVAB (ASVAB, 2023). Subscore information displayed under the ASVAB tests section is a hybrid of Figure 2.1b and c, showing percentile ranks for different norming groups and a visual presentation of the subscores in a profile (score bands) format to assist with interpreting relative differences. Unlike the excerpt from the ABIM MOC report, the subscore profiles in this sample have far less overlap, likely reflecting greater reliability (e.g., narrower bands) and validity (variability in performance), although it is possible that the bands were constructed differently, based on a lower standard error criterion or using a different standard deviation, which could be used to distort the results and create the perception of greater reliability and validity. The recipient of this report can see that they performed considerably better in Word Knowledge than in Electronics Information. When considering the purpose of the ASVAB and that the subscores are used to determine entrance eligibility, this level of detail is important in placement decisions for various career positions and branches of service.

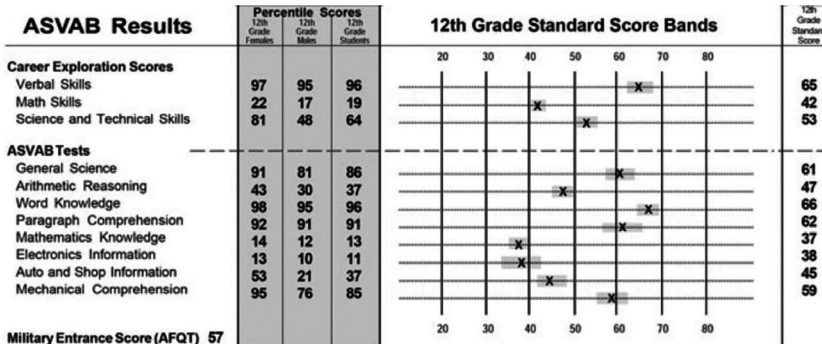


Figure 2.6 Excerpt from an ASVAB score report.

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

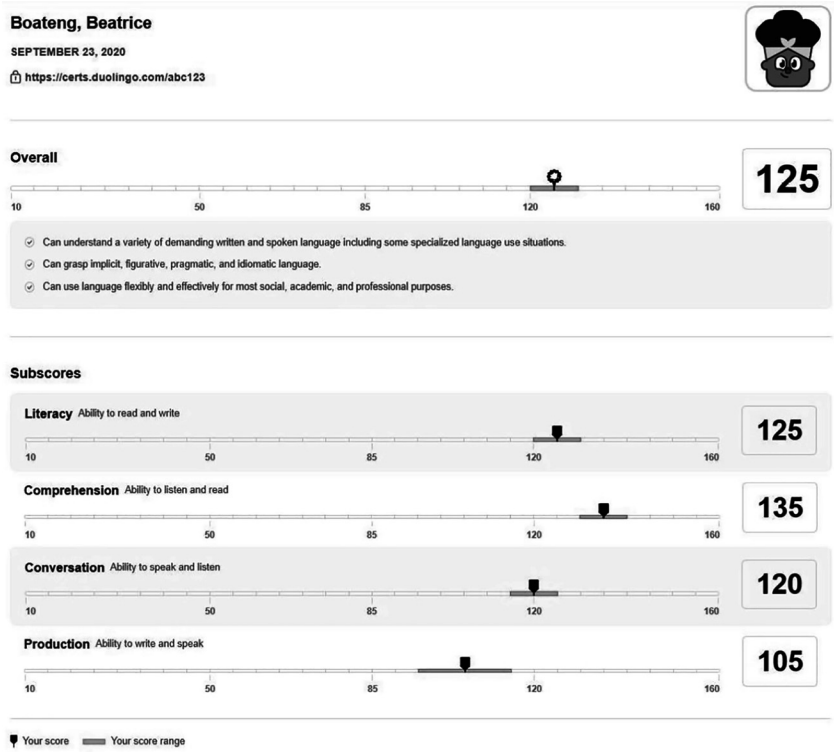


Figure 2.7 Excerpt from the sample Duolingo English Test Certificate. Reproduced with permission. All rights reserved.

2.4.6 Duolingo English Test

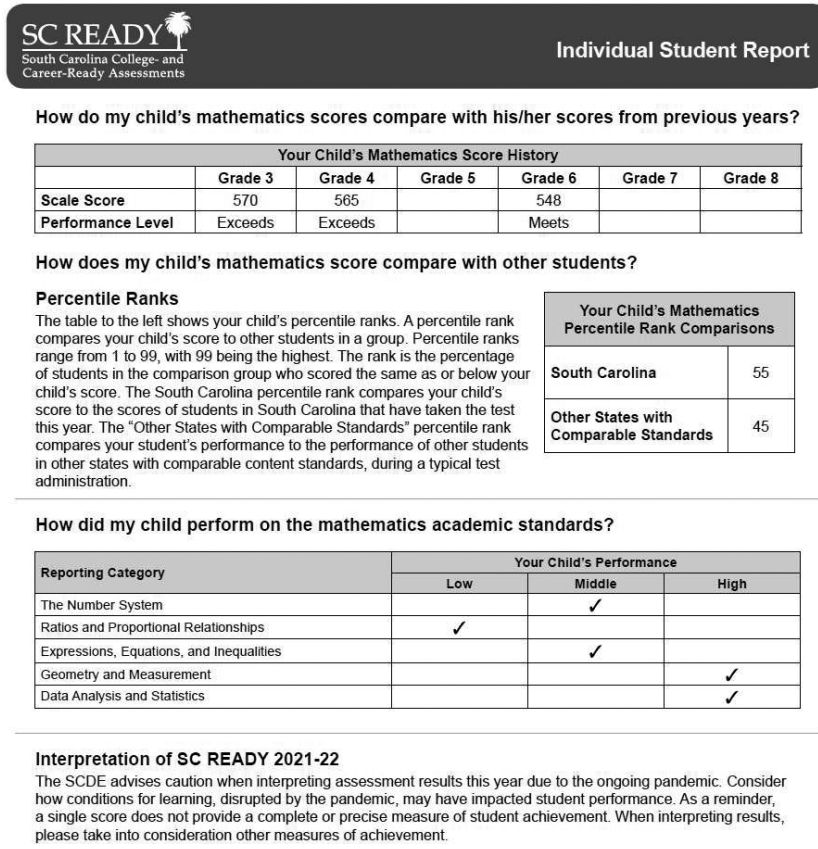
The Duolingo English Test is a summative assessment designed to inform undergraduate and graduate admission decisions related to general English language proficiency. Figure 2.7 presents an excerpt from a sample certificate (score report) test takers would receive after completing the test (Duolingo, 2023). Similar to how subscores are reported on the ASVAB report, students receive both their numeric score and a score range on a comparable scale across all subscores. In this example, due to the nonoverlapping score ranges (profiles), the student can see that they performed relatively better in Comprehension and relatively worse in Production. The wider score range for Production likely suggests less precision, and possibly fewer items, compared to the other content areas. Also similar to the ASVAB, the lack of explanatory text on this certificate makes it difficult to know exactly how the score ranges were computed and whether the separation between ranges reflects greater

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

reliability and validity. However, though not mentioned directly on the certificate, the Duolingo website does include a research report (LaFlair, 2020) that describes the psychometric properties of the subscores and evidence for how they can be used. A comprehensive review of recommended approaches for assessing subscore value will be covered in Chapter 3.

2.4.7 South Carolina College- and Career-Ready Assessments (SC READY)

The South Carolina College- and Career-Ready Assessments (SC READY) are formative assessments designed to measure the extent to which students are on track for the next grade and ultimately college and career readiness based



SC READY
South Carolina College- and Career-Ready Assessments

Individual Student Report

How do my child's mathematics scores compare with his/her scores from previous years?

Your Child's Mathematics Score History						
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Scale Score	570	565		548		
Performance Level	Exceeds	Exceeds		Meets		

How does my child's mathematics score compare with other students?

Percentile Ranks

The table to the left shows your child's percentile ranks. A percentile rank compares your child's score to other students in a group. Percentile ranks range from 1 to 99, with 99 being the highest. The rank is the percentage of students in the comparison group who scored the same as or below your child's score. The South Carolina percentile rank compares your child's score to the scores of students in South Carolina that have taken the test this year. The "Other States with Comparable Standards" percentile rank compares your student's performance to the performance of other students in other states with comparable content standards, during a typical test administration.

Your Child's Mathematics Percentile Rank Comparisons	
South Carolina	55
Other States with Comparable Standards	45

How did my child perform on the mathematics academic standards?

Reporting Category	Your Child's Performance		
	Low	Middle	High
The Number System		✓	
Ratios and Proportional Relationships	✓		
Expressions, Equations, and Inequalities		✓	
Geometry and Measurement			✓
Data Analysis and Statistics			✓

Interpretation of SC READY 2021-22

The SCDE advises caution when interpreting assessment results this year due to the ongoing pandemic. Consider how conditions for learning, disrupted by the pandemic, may have impacted student performance. As a reminder, a single score does not provide a complete or precise measure of student achievement. When interpreting results, please take into consideration other measures of achievement.

Figure 2.8 Excerpt from the fourth page of the sample SC Ready Individual Student Report. Reproduced with permission. All rights reserved.

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

on state-defined standards set by the South Carolina Department of Education (SC READY, 2023). Students complete versions of this assessment throughout their elementary and secondary education in English language arts (ELA), mathematics, science, and social studies. Figure 2.8 presents an excerpt from a sample sixth-grade student score report that displays categorical subscores relative to meeting academic standards in mathematics. This report would be provided to students and their parents as well as the school to inform curriculum improvements and identify areas where students may need additional support. Additionally, score information is aggregated across schools and districts for the South Carolina state-level summary of results. In the excerpt, the student would see that their overall mathematics performance met sixth-grade expectations, but it also reflects a decrease from their performance in third and fourth grades. Exploring the categorical subscores suggests that their performance relative to the state's academic standards was low in Ratios and Proportional Relationships, which may explain the decrease in overall mathematics performance and represent a key area for remediation.

2.4.8 Comprehensive Clinical Science Examination (CCSE)

The Comprehensive Clinical Science Examination® (CCSE) is designed to assess learning progress and readiness for medical students to take the United States Medical Licensing Examination® (USMLE) Step 2 Clinical Knowledge Examination (CCSE, 2023). Figure 2.9 presents an excerpt from the third page of a sample CCSE score report in which subscores are presented in a categorical format as either lower, same, or higher relative to a comparison group of medical students enrolled at accredited schools testing for the first time. Additionally, the report provides the student's equated percent correct score for each content area and the average for the comparison group. Interpretive text on the prior page explains that the equated percent correct scores "may be slightly lower or higher than the actual percentage of questions you answered correctly on this specific exam form because they are statistically adjusted to account for slight variations in exam form difficulty."

The score report also includes the percent of questions per content area, which reveals substantial differences in weighting between categories. This may be useful to score users in helping interpret the results – lower performance in a category with fewer items may be worth less remedial effort than average performance in a category with many items. Additionally, it can be deduced by the weights that items are coded multiple times across the system, discipline, and physician task (not shown) dimensions. This is likely the result of creating realistic test questions but could also have implications where the overlap causes the subscores to become more redundant (Feinberg & Wainer, 2014).

SUBJECT EXAMINATION PROGRAM

COMPREHENSIVE CLINICAL SCIENCE EXAMINATION (CCSE)

EXAMINEE PERFORMANCE REPORT

NBME

ID: 000000000
Name: Student A
000000 - Generic Medical School

Total CCSE Score: 230
Test Date: Month Day, Year

	Your EPC Score	Comparison Group Average EPC Score	Score Comparison: Lower	Same	Higher	% of Questions
Performance by System						
Legal/Ethical, Professionalism, System-based Practice, Pt Safety	**87**	81		■		10 - 14%
Cardiovascular System	**73**	72		■		8 - 10%
Immune System & Blood & Lymphoreticular Systems	**63**	78	■			7 - 11%
Gastrointestinal System	**63**	75	■			7 - 9%
Respiratory System	**60**	72		■		7 - 9%
Musculoskeletal Sys/Skin & Subcutaneous Tissue	**70**	71		■		6 - 10%
Behavioral Health	**55**	80	■			6 - 8%
Nervous System & Special Senses	**63**	74		■		6 - 8%
Endocrine System	**65**	78		■		4 - 6%
Female Reproductive & Breast	**66**	75		■		4 - 6%
Multisystem Processes & Disorders	**57**	74	■			4 - 6%
Pregnancy, Childbirth & the Puerperium	**83**	75		■		4 - 6%
Renal & Urinary System & Male Reproductive	**74**	77		■		4 - 6%
Performance by Discipline						
Medicine	**66**	75	■			50 - 60%
Surgery	**68**	74	■			25 - 30%
Pediatrics	**73**	76		■		20 - 25%
Obstetrics & Gynecology	**80**	77		■		10 - 20%
Psychiatry	**64**	81	■			10 - 15%

Figure 2.9 Excerpt from the third page of the CCSE Examinee Performance Report. Copyright © 2022 by the National Board of Medical Examiners (NBME). Reproduced with permission. All rights reserved.

2.4.9 United States Medical Licensing Examination (USMLE) Annual School Report

The United States Medical Licensing Examination (USMLE) is a three-step exam sequence to obtain medical licensure in the United States. In addition to individual score reports to medical students, annual reports are provided to schools meeting certain criteria (e.g., at least 20 students) to facilitate comparisons within and between institutions (USMLE, 2023). Figure 2.10

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

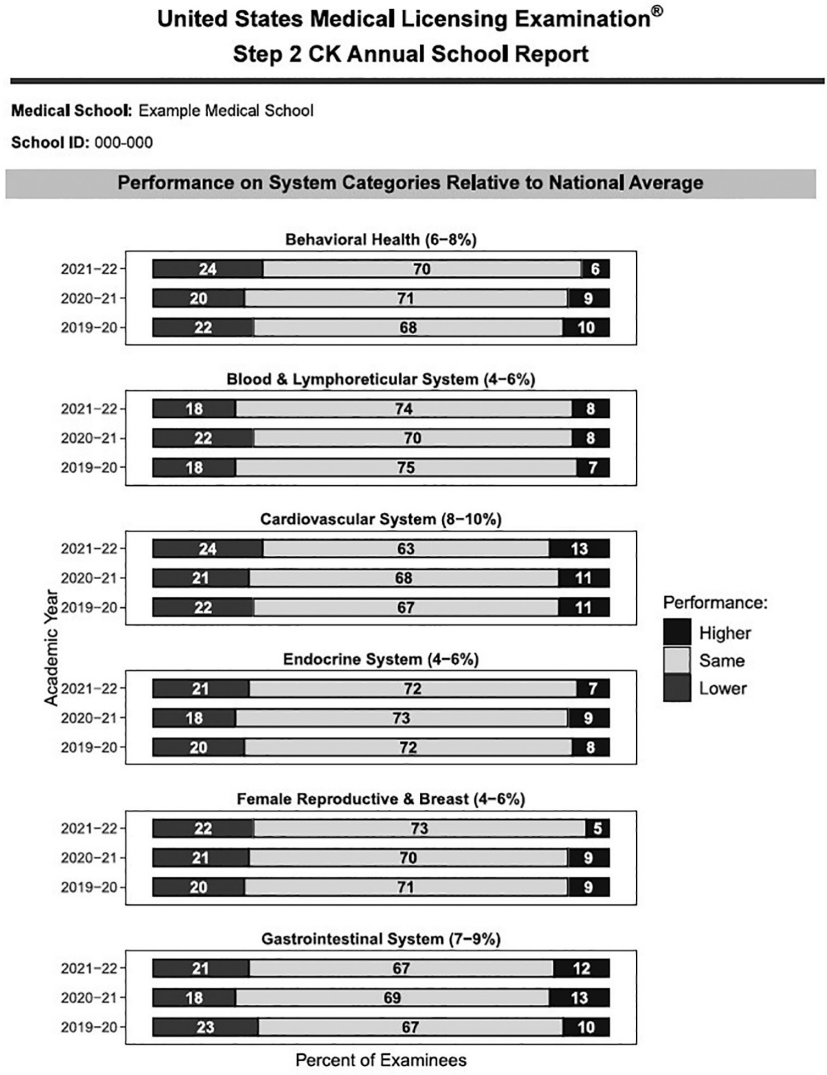


Figure 2.10 Excerpt from the seventh page of the USMLE Step 2 CK School Report. Copyright © 2022 by the Federation of State Medical Boards (FSMB) and National Board of Medical Examiners (NBME). Reproduced with permission. All rights reserved.

presents an excerpt of the subscore performance summary from a sample annual school report on the Step 2 Clinical Knowledge (CK) examination. For each content area, the percent of students from the particular school that

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

performed higher than, similar to, or lower than a national comparison group are displayed as well as what those percentages were for the previous two years. This reporting style supports two inferences – by comparing against the national average, users can interpret the between-school normative context, and with the prior years' information, users can also interpret the within-school historical context. The similarity in results across categories is likely a sign that the subscores are highly correlated. Additionally, category weights are clearly displayed to convey the relative importance and precision of the underlying scores.

2.4.10 Critical Thinking Assessment

Figure 2.11 presents an institution score report excerpt from a postsecondary education critical thinking assessment designed to assess a common set of skills and competencies deemed relevant to learning outcomes for college graduates. In this case, the lighter wider profile band is expressed as a box plot of average scores for all institutions in the comparison group, with the box representing the middle 50% of institution average scores and the

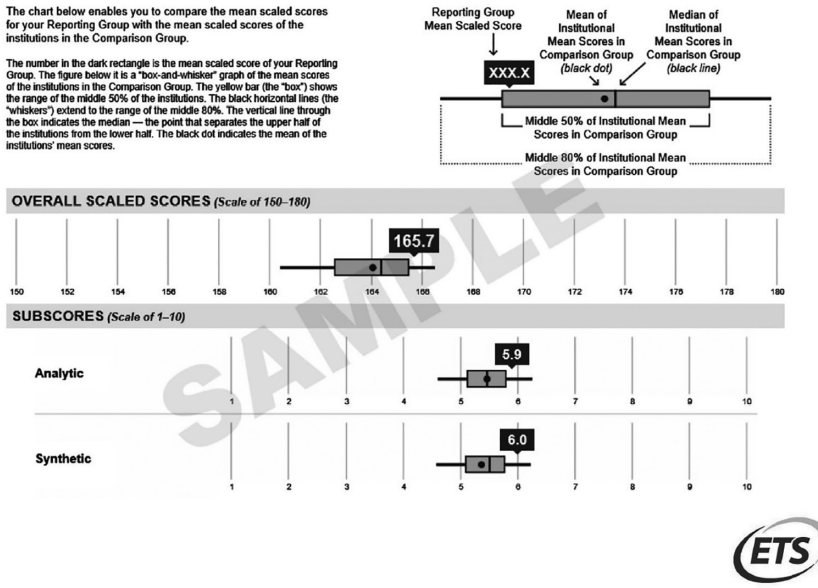


Figure 2.11 Excerpt from an institutional report for an assessment measuring critical thinking. Copyright © 2023 by Educational Testing Service (ETS), www.ets.org. Reproduced with permission. All rights reserved.

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

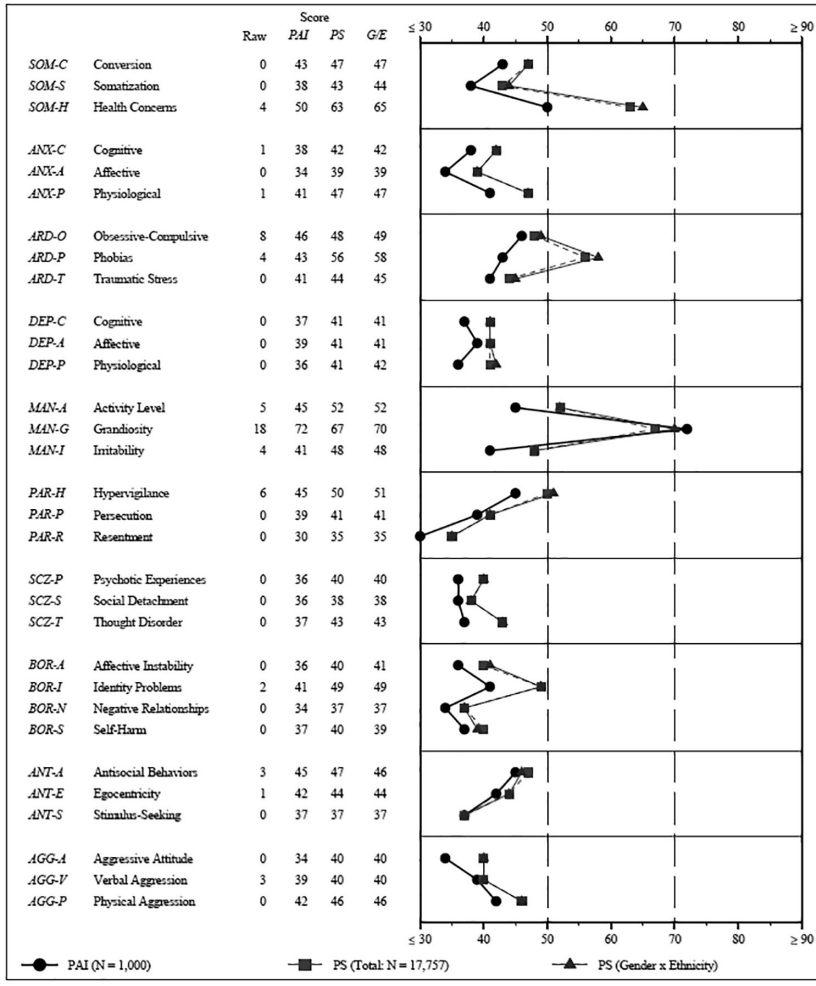
middle 80% within the whisker boundaries. The solid circle and vertical line inside the box indicate the grand mean and median, respectively. The darker narrower band represents the average student performance from this particular institution. Using this information, the recipient can infer that the participating students from their institution performed above average in both the analytic and synthetic subscores. Not shown here, other elements in the full report provide contextual information on what overall score ranges fit with developing, proficient, or advanced proficiency levels. However, subscore ranges are not mapped to these descriptions, so it would be difficult for the recipient to know the extent to which any of the subscores met expectations.

2.4.11 Law Enforcement, Corrections, and Public Safety Personality Assessment Inventory® (PAI)

The Personality Assessment Inventory (PAI),[1] first published in 1990, is a 344-item self-report measure of personality and psychopathology that includes 22 scales, 10 of which contain subscores (Morey, 2004). The PAI Law Enforcement, Corrections, and Public Safety Selection Report was specially designed by licensed psychologists to assess emotional stability during the selection process and is based on a normed sample of 18,000 public safety job applicants (PAI, 2023). Figure 2.12 presents an excerpt of an applicant's subscore profile relative to job applicants (*PS*), the general community (*PAI*), and a matching ethnic-gender group, which in this case was Caucasian males. Performance for each subscore is represented by a *T* score, which is scaled to have a mean of 50 and standard deviation of 10, normed within the respective comparison group. Raw scores are also provided, but the wide variability likely indicates substantial differences in the number of items, making them difficult to interpret. The profile reveals that the sample job applicant is similar across most subscales for all comparison groups but did demonstrate elevated levels of grandiosity and, relative to other job applicants and the specific ethnic-gender comparison groups, slightly higher levels of health concerns and phobias than would be considered typical for the general community. The evaluating psychologist could use this information to follow up on whether advancing the candidate presents a risk or if additional measures are needed to further explore potential concerns.

[1] The PAI is a copyrighted instrument and may not be used or reproduced in whole or in part, in any form or language, or by any means without written permission of PAR (www.parinc.com).

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press



Note. This profile is based on calculations from all applicants in the screening process, regardless of the final selection decision. Ethnic group used for gender- by ethnic-specific profile is Caucasian Male. Refer to the Professional Issues chapter of the manual for the gender by ethnic group sample sizes.

Figure 2.12 Excerpt from the PAI Law Enforcement, Corrections, and Public Safety Selection Report. Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc. (PAR), 16204 North Florida Avenue, Lutz, Florida 33549 from the Personality Assessment Inventory by Leslie C. Morey, PhD, Copyright 1991. Further reproduction is prohibited without permission of PAR.

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

2.4.12 Assessment Tools for Teaching and Learning (asTTIe)

New Zealand's Assessment Tools for Teaching and Learning (asTTle) is a formative assessment system administered by the country's Ministry of Education and provides diagnostic performance information to support student learning outcomes and improve instruction (Brown, O'Leary, & Hattie, 2019). The assessment system is not intended to be used for evaluation or punitive purposes – only for informing improvement. Figure 2.13 presents an excerpt from an asTTle console report that is an interactive online score dashboard. Teachers use the subscore information, which in this format is presented as gauge plots and boxplots to explore performance for a cluster of students or drill into performance for a particular student. In the figure, for instance, the user could infer that the students from New Zealand schools are demonstrating lower knowledge and understanding compared to other students at the same year level, but they do have a positive attitude. The gauges without any information could indicate that those characteristics were not assessed or perhaps reflect extreme poor performance, which is impossible to know without additional explanatory text.

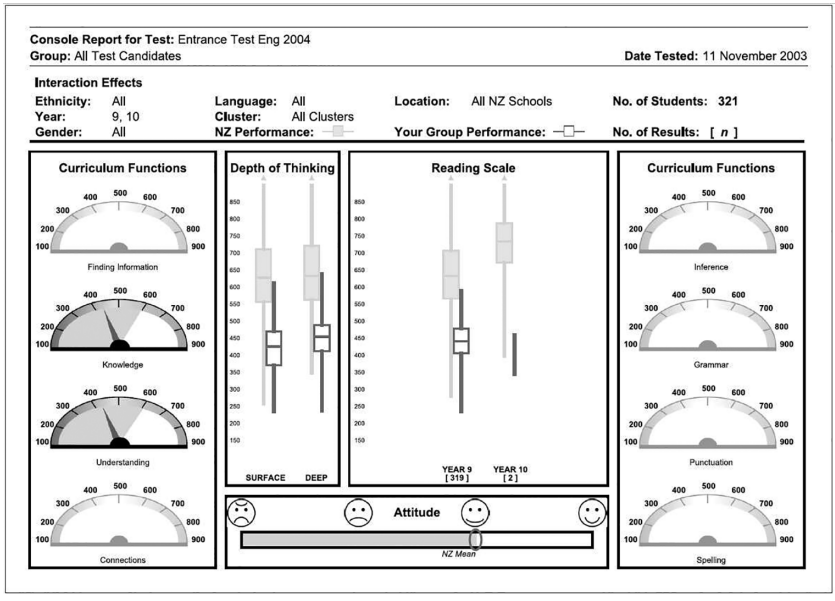


Figure 2.13 Excerpt of an asTTle console report.
From Brown, O'Leary, and Hattie (2019). Reproduced with permission. All rights reserved.

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

2.5 Generalizing Score Reporting Best Practices to Subscore Reporting Best Practices

The decision on what to report is a function of the intended inferences, using data displays that support those inferences at an appropriate level of granularity (*psychometric quality*) and presented in a way that users can correctly interpret those inferences (*type of users*). If they follow current best practices in the assessment industry, then the test publisher knows the intended inferences and who the users of the results are, as those details would have been clearly defined in the initial test design. A test's design, sometimes referred to as test specifications, articulates all critical aspects of the assessment life cycle to align the purpose of the test and eligibility requirements with the desired inferences carried through to score reporting (Lane et al., 2015).

However, how does the test publisher know users are making the correct inferences from the score report? How does the test publisher know if users desire additional unsupported inferences or are not finding utility in certain reported results? If so, why is there a disconnect? Are there differences in any of these questions between types of users, by either construct-relevant proficiencies or perhaps construct-irrelevant factors like gender, race, native language, or disabilities? Additionally, the information to be presented and the needs of score users may change over time as other relevant factors evolve such as shifts in practice and curriculum, technology enhancements, the composition of the examinee population, and competition from rival testing organizations. This can be particularly troublesome for subscores as users increasingly demand more detailed feedback for diagnostic purposes that were never part of the original test design, which ultimately leads to a desire to receive subscores that lack sufficient psychometric properties.

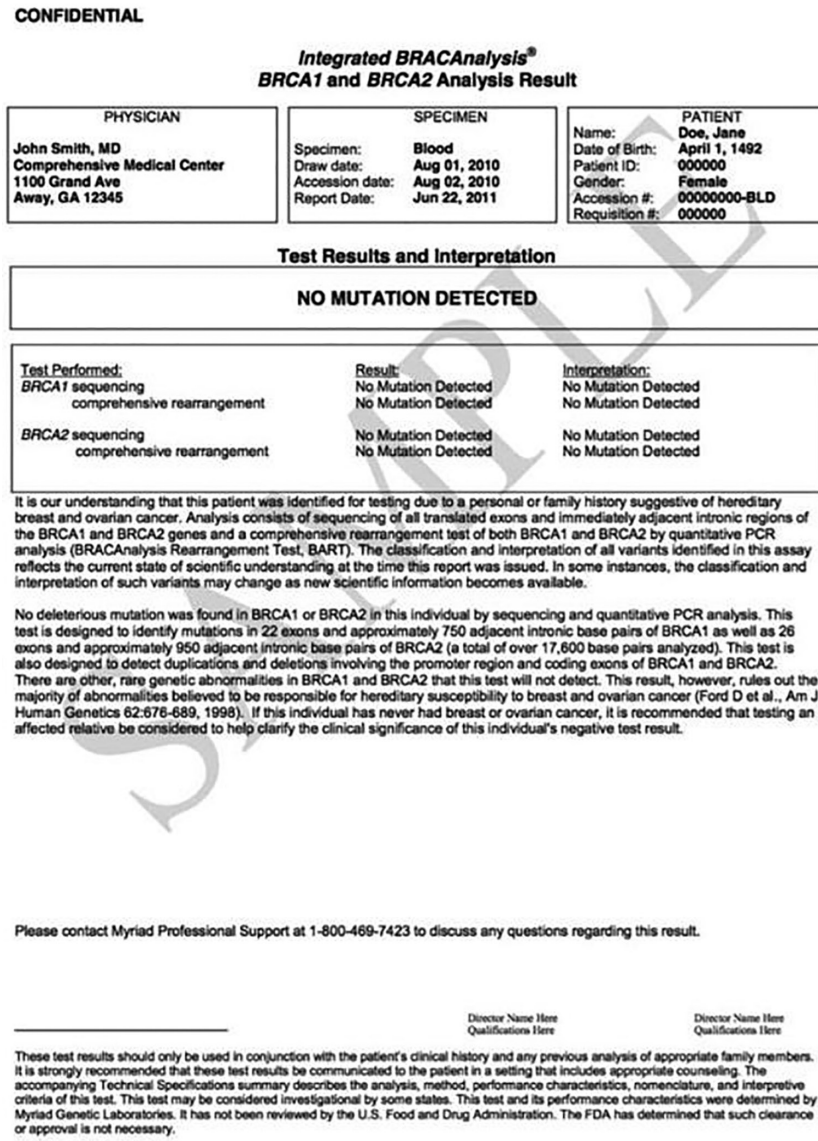
These concerns can be addressed by a systematic approach to score report development. Several researchers have proposed similar iterative and multistep frameworks for score report design that include gathering information on user needs and user characteristics pertinent to reporting, creating preliminary mockups, engaging with stakeholders to collect feedback (e.g., focus groups), refining a prototype, and then finalizing the score report template for operational deployment (Zapata-Rivera, VanWinkle, & Zwick, 2012; Zenisky & Hambleton, 2012; Hambleton & Zenisky, 2013; Zenisky & Hambleton, 2015; Slater, Livingston, & Silver, 2019). Utilizing a comprehensive framework in score report design not only helps a test publisher understand user needs and optimize alignment with a test's purpose but can also reveal concerns that may be overlooked from a measurement perspective.

Previous research has suggested general principles for displaying data in either a table (Feinberg & Wainer, 2011) or a graph (Tufte, 2001; Friendly & Wainer, 2021), though research specifically on displaying information on score reports suggests there is no such thing as a "best visualization," as it depends on many factors, including the purpose of the display, prior knowledge, information needs, and characteristics of the user (Hegarty, 2019). There are, of course, many ways to poorly display data, which include showing too little or too much information, emphasizing the trivial, skewing the axes on a graph, and presenting needless decimals (Wainer, 1984). Another important consideration is including links to relevant documentation concerning the reliability and validity of the subscores. Some test publishers include this information in a technical report or manual on their website and certainly when required, such as for state and federal assessments like the New York State Testing Program (NYSTP, 2023). Some of the sample score reports mentioned in this chapter included generic links to the publisher's website, but none provided a clear path to this technical information, and many private test publishers do not make this information publicly available.

A systematic approach to score report development can explore how the target recipients perceive the results, exploring questions such as how much detail is too little or too much, interactions with working memory and capacity for attention, and how best to organize information to facilitate comprehension. Additionally, there is an emotional component to interpreting a score report. For instance, a failing grade on a high-stakes test often can have several implications beyond needing to retest – embarrassment among the test taker's friends and family, financial implications of a delay or stop in a career path, and managing the psychological disconnect if the results are vastly different from what was expected. It may not be the job of the test publisher to provide a solution for these concerns, but being aware of the holistic interpretation can help test publishers be more sensitive in communicating disappointing information and present themselves as an advocate.

A serious example of empathy in communicating results described by Wainer (2015) involves the genetic reports a patient would receive when waiting to find out if they carry one of several mutated genes that substantially increase their risk of developing breast cancer. Figure 2.14 presents a current version of the report that provides ample detail and description but obfuscates the key finding – that no mutation was detected. Figure 2.15 presents a suggested redesign, where the important finding is now much more obvious. However, if a mutation had been detected, then the recipient may prefer the current version (Figure 2.14) with plenty of explanatory text and a more

sensitive presentation. It may be challenging from an operational perspective, but with different combinations of results and types of recipients, more than one score reporting template may improve how results are being received.



CONFIDENTIAL

Integrated BRACAnalysis®
***BRCA1* and *BRCA2* Analysis Result**

PHYSICIAN	SPECIMEN	PATIENT
John Smith, MD **Comprehensive Medical Center** **1100 Grand Ave** **Away, GA 12345**	Specimen: **Blood** Draw date: **Aug 01, 2010** Accession date: **Aug 02, 2010** Report Date: **Jun 22, 2011**	Name: **Doe, Jane** Date of Birth: **April 1, 1492** Patient ID: **000000** Gender: **Female** Accession #: **00000000-BLD** Requisition #: **000000**

Test Results and Interpretation

NO MUTATION DETECTED

Test Performed:	Result:	Interpretation:
BRCA1 sequencing	No Mutation Detected	No Mutation Detected
comprehensive rearrangement	No Mutation Detected	No Mutation Detected
BRCA2 sequencing	No Mutation Detected	No Mutation Detected
comprehensive rearrangement	No Mutation Detected	No Mutation Detected

It is our understanding that this patient was identified for testing due to a personal or family history suggestive of hereditary breast and ovarian cancer. Analysis consists of sequencing of all translated exons and immediately adjacent intronic regions of the BRCA1 and BRCA2 genes and a comprehensive rearrangement test of both BRCA1 and BRCA2 by quantitative PCR analysis (BRACAnalysis Rearrangement Test, BART). The classification and interpretation of all variants identified in this assay reflects the current state of scientific understanding at the time this report was issued. In some instances, the classification and interpretation of such variants may change as new scientific information becomes available.

No deleterious mutation was found in BRCA1 or BRCA2 in this individual by sequencing and quantitative PCR analysis. This test is designed to identify mutations in 22 exons and approximately 750 adjacent intronic base pairs of BRCA1 as well as 26 exons and approximately 950 adjacent intronic base pairs of BRCA2 (a total of over 17,600 base pairs analyzed). This test is also designed to detect duplications and deletions involving the promoter region and coding exons of BRCA1 and BRCA2. There are other, rare genetic abnormalities in BRCA1 and BRCA2 that this test will not detect. This result, however, rules out the majority of abnormalities believed to be responsible for hereditary susceptibility to breast and ovarian cancer (Ford D et al., Am J Human Genetics 62:676-689, 1998). If this individual has never had breast or ovarian cancer, it is recommended that testing an affected relative be considered to help clarify the clinical significance of this individual's negative test result.

SAMPLE

Please contact Myriad Professional Support at 1-800-469-7423 to discuss any questions regarding this result.

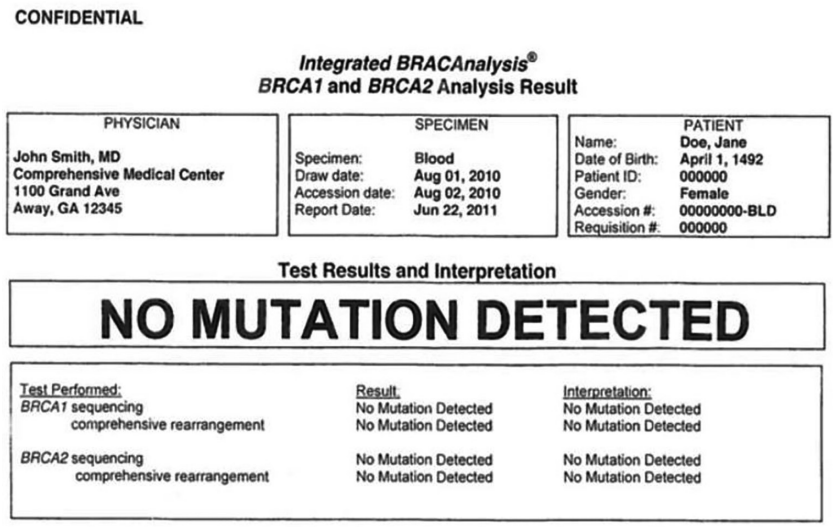
Director Name Here
Qualifications Here

Director Name Here
Qualifications Here

These test results should only be used in conjunction with the patient's clinical history and any previous analysis of appropriate family members. It is strongly recommended that these test results be communicated to the patient in a setting that includes appropriate counseling. The accompanying Technical Specifications summary describes the analysis, method, performance characteristics, nomenclature, and interpretive criteria of this test. This test may be considered investigational by some states. This test and its performance characteristics were determined by Myriad Genetic Laboratories. It has not been reviewed by the U.S. Food and Drug Administration. The FDA has determined that such clearance or approval is not necessary.

Figure 2.14 Notice of negative finding of gene mutation from Myriad Laboratories. *Source:* Wainer (2015).

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press



CONFIDENTIAL

Integrated BRACAnalysis®
BRCA1 and *BRCA2* Analysis Result

PHYSICIAN	SPECIMEN	PATIENT
John Smith, MD	Specimen: Blood	Name: Doe, Jane
Comprehensive Medical Center	Draw date: Aug 01, 2010	Date of Birth: April 1, 1492
1100 Grand Ave	Accession date: Aug 02, 2010	Patient ID: 000000
Away, GA 12345	Report Date: Jun 22, 2011	Gender: Female
		Accession #: 00000000-BLD
		Requisition #: 000000

Test Results and Interpretation

NO MUTATION DETECTED

Test Performed:	Result:	Interpretation:
BRCA1 sequencing	No Mutation Detected	No Mutation Detected
comprehensive rearrangement	No Mutation Detected	No Mutation Detected
BRCA2 sequencing	No Mutation Detected	No Mutation Detected
comprehensive rearrangement	No Mutation Detected	No Mutation Detected

Figure 2.15 Suggested revision of notice of negative finding of gene mutation that emphasizes the principal message.
Source: Wainer (2015).

2.6 Concluding Remarks

The decision of how subscores should be reported is complex and depends on many factors. Most important are the prospective uses of the subscores, their psychometric properties, the type of test and associated purpose, and the stakeholders to whom the results would be reported. For summative assessments, reporting subscores is particularly challenging, as they tend to lack sufficient psychometric properties (Sinharay, 2010) – more on that in the next chapters. However, this is not a surprising finding, given that constructing a test primarily designed to support inferences from a single total score, which is typically the case in summative assessments, leads to including items that correlate highly with the overall score because they are deemed essential to maximize reliability. The items that have lower correlations with the overall score, because they measure a slightly different construct and have the potential to contribute more toward subscore value, are likely to be systematically excluded. Thus, efforts to improve subscore distinctiveness (validity) would in turn diminish test reliability and the utility of the total score and, hence, detract from the primary purpose of the assessment. As stated by Brennan (2012), "If test scores fit a unidimensional model, a psychometrically compelling argument cannot

https://doi.org/10.1017/9781009413701.003 Published online by Cambridge University Press

be mounted for reporting any subscores since, by definition, there is only one proficiency or latent trait." Thus, tests that are designed for summative assessment will almost always struggle to have meaningful subscores to report, as that would inevitably confound the primary purpose.

Further, when subscores in a raw score format do have value, they should be reported numerically only after being converted to an established scale. Observed differences between raw subscores either within the same score report or compared longitudinally against a prior attempt, may be confounded due to differences in how the underlying subtests were constructed. Test publishers may be tempted to make the subscore scales comparable to the total score scale (e.g., same mean and standard deviation) or equal to a fraction of the total score scale such that the sum of all subscores equals the total. These choices often contribute to misinterpretation, given that subscores typically have lower reliability than the total score. Therefore promoting comparisons using the same scale can be problematic as subscores are expected to contain more error than the total score and their summation would include compounding error. In addition, reported subscores should be equated or at least linked so that the definition of strong performance in one content area does not change between forms of a test within a single administration or over time. In typical cases, equating is feasible for the total score but not for subscores due to the small number of items. Some work on equating subscores has been done by Puhan and Liang (2011).

Given these challenges, other than not reporting subscores at all, the potential to mislead users can be mitigated with less granular reporting styles and appropriate interpretive language, which can be determined only by engaging with a representative group of users and assessing the extent to which they correctly interpret the score information. Consider how Daisy's story would have been different had the presentation of her results been organized in a more coherent report at a valid and defensible level of detail. Perhaps her parents would have continued pursuing their desired outcome, but at least they wouldn't have perceived evidence from an official score report that could be weaponized to pressure the school.

However, before score users can be engaged to consider and help evaluate report design, the test publisher needs to first understand the psychometric properties of the subscores. Knowledge of the subscore statistical characteristics can inform realistic conversations about their limitations and appropriate next steps on what is defensible to report, if anything, or what could be done to improve their value, if anything. How exactly can a test publisher evaluate subscore quality? How much is good enough? That is the focus of our next chapter.