Implementation of Evidence-Based Standard Setting for Advanced Placement Exams

About Advanced Placement Exams

The Advanced Placement® (AP®) Program allows motivated and academically prepared students to take college-level courses in high school, and it gives them the chance to earn college credit, advanced placement, or both. AP courses, offered in 40 subjects, end with a rigorous exam that helps students develop critical thinking, solid argumentation, and a broad perspective—skills essential for college and beyond. Taking AP courses signals to college admissions officers that students have pursued the most challenging curriculum available. Research shows that students scoring 3 or higher on AP Exams often achieve greater academic success in college and are more likely to earn a degree than students who haven't taken AP (Beard et al., 2019). Each AP teacher's syllabus is reviewed by top college faculty, and AP Exams are created and scored by college faculty and experienced AP teachers. Most U.S. four-year colleges and universities, along with over 3,300 institutions worldwide, grant credit, advanced placement, or both based on successful AP Exam scores.

Evidence-Based Standard Setting for Advanced Placement Exams

Performance standards provide the foundation for determining appropriate exam cut scores and validating score interpretations for Advanced Placement Exams. AP Exam scores and course grade correlates are defined by the AP Program to provide a representation of test takers' expected achievement, or performance standard, in the comparable college course (see Table 1). College Board has recently adopted a comprehensive, evidence-based approach to choosing cut scores that align with the predetermined performance standards, known as Evidence-Based Standard Setting (EBSS; McClarty et al., 2013). The following sections highlight traditional practices for standard setting and explore how these practices were enhanced using EBSS to set AP cut scores.

Table 1. AP and College Course Grade Correlates

College Course Grade Equivalent	AP Exam Score	Credit Recommendation
A+, A	*5*	*Extremely well qualified*
A-, B+, B	*4*	*Well qualified*
B-, C+, C	*3*	*Qualified*
-	*2*	*Possibly qualified*
-	*1*	*No recommendation*

Standard Setting

Performance standards for testing programs are often related to content standards, normative standards, and/or criterion standards. Content standards define performance based on mastery of content and skills. Normative standards ensure AP Exam scores align with college grading expectations and research recommendations. Criterion/predictive standards assess how AP Exam performance aligns with subsequent college performance. Traditionally, AP standard setting focused on content-based methods, and relatively small groups of subject-matter experts (SMEs) were required to judge which score points relate to the skills and knowledge defined by performance standards. There was little attention to norm-referenced or predictive information regarding hypothesized outcomes about the test takers' performance. We now have greater capacity to conduct large-scale studies and analyze data for selecting normative-, predictive-, and content-based cut scores, thus enhancing the validity evidence for AP standard setting.

Rationale for and implementation of EBSS in AP

The AP Program employs the Evidence-Based Standard Setting (EBSS) framework to systematically collect and analyze validity evidence, understand the relationship between the AP Exam and college outcomes, and establish performance standards that address normative, content, and predictive goals. EBSS involves compiling data from various research studies for each subject, which enables experts to make evidence-based decisions about exam cut scores with substantial empirical support. EBSS research studies include surveys of higher education instructors with questions about the AP Exams and college student performance. A compilation of EBSS aggregated data is developed for each AP subject at the time of standard setting. This compilation includes a summary of input from college instructors on course content, exam difficulty, and performance expectations. It summarizes historical test taker performance, comparisons of college performance between students who did and didn't take AP, and studies where college students take the AP Exam to compare their performance with course grades. Subject matter experts and policymakers review these aggregate-level compilations. They balance content, normative, and predictive standards to evaluate and decide on appropriate cut scores representing different performance levels.

Evidence collected for EBSS

The compilation of EBSS data for a given AP subject includes the following types of aggregated data as evidence to support the cut score decisions:

Data Compiled from the AP Exam Administration

Test taker performance characteristics summarize the relationship between AP Exam and SAT® Suite of Assessments performance (accounting for typical age at time of testing for each). Test takers' performance on the SAT Suite gives context because the SAT Suite also measures college readiness in terms of academic reading, writing, and math skills, and the correlations between SAT and AP scores are strong. Detailed test taker performance on types of questions at

various potential cut scores are also considered, in addition to performance across other exams. For example, for similar AP subjects, such as calculus-related exams, the relationship between performance on the related AP Exams is also described. Name, race, ethnicity, gender, or any other test taker characteristic beyond their AP and SAT scores isn't included.

Score distributions and exam statistics, including the difficulty of various exam components in terms of historical and recent test taker performance, provide context for the standard setting and highlight factors that might influence cut score decisions.

Impact data describe historical AP Exam score distributions and the AP Exam score distributions based on possible cut scores in the standard setting study.

Special Studies to Gather Additional Data and Evidence for Standard Setting

Longitudinal evidence of student performance for students who took AP and those that didn't take AP comes from studies comparing the performance of both groups of students in subsequent college courses. These studies aim to connect college and AP student performance. Longitudinal performance studies require college performance data for a subsequent course following the introductory college course comparable to the AP subject. At the time of the standard setting, this longitudinal evidence is based on the prior AP standards. Results based on updated standards aren't available until several years after the standard setting study.

Higher education participant information provides demographics for the instructors and institutions recruited and included in EBSS for a particular subject who teach an introductory college course comparable to AP. Data are typically collected from at least 100 instructors from institutions of varied sizes and types, such as two-year and four-year, and with varying admission rates, in over two-thirds of the states. The data represents 3,000 to more than 25,000 students per AP subject. Note that these EBSS studies include at least five times as many instructors as the previous AP standard-setting studies, depending on the subject. Characteristics of the courses and students these instructors teach, and for which they're reporting information as a part of EBSS, are also included.

Course comparisons provide the alignment of college course content submitted by the higher education instructors with AP course descriptions, and they include information about their curriculum, assessments, and grading policies. Course grade distributions report college students' grades achieved in comparable college exams and courses as reported by the EBSS participants. The instructors provide final course and exam grade distributions for their courses over recent years.

Multiple-choice and free-response question (MCQ and FRQ) evaluations provide the results from the higher education instructors examining specific AP Exam questions and identifying the expected lowest grade level of a college student (e.g., A, B) likely to answer particular questions correctly.

Percent ratings reflect the instructors' expectations of how their college students would perform on the AP Exam as a whole. Instructors evaluate an AP Exam in relation to a comparable college course. They determine the percentage of score points needed to achieve a grade level (e.g., A, B.) and the percentage of students achieving those grades. A college comparability study is occasionally conducted to directly examine the relationship between course grades and AP Exam scores. A sample of college students takes the AP Exam, and their exam scores are compared to their grades in the comparable college course in which they're enrolled.

Following the data collection and analysis, College Board staff review the compiled EBSS results to determine the final cut scores. They take into account all data inputs, and they evaluate average performance on MCQ and/or FRQ sections of the exam in relation to the level of expected knowledge and skills.

In conclusion, EBSS provides a comprehensive framework for selecting cut scores that match performance standards for AP Exams. The framework integrates multiple sources of validity evidence while taking into consideration content standards, normative standards, and criterion standards. This approach ensures that cut scores are based on a robust foundation of research and expert judgment, which ultimately enhances the validity of AP Exam score interpretations.

References

Beard, J. J., Hsu, J., Ewing, M., & Godfrey, K. E. (2019). Studying the relationships between the number of APs, AP performance, and college outcomes. Educational Measurement: Issues and Practice, 38(4), 42–54.

McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-based standard setting: Establishing a validity framework for cut scores. Educational Researcher, 42(2), 78–88.