Program Critique: Standard Setting Procedures and Performance Level Descriptors for Assessment Programs Designed to Identify and Measure Giftedness

Lizzie Kahle
Professor Mike Russel
Boston College

Assessment Programs MESA 7240
December 12, 2023

Part One: An Overview of PLDs and Standard Setting Procedures

In the realm of educational testing, an assessment program is a structured process designed to measure the knowledge, skills, and abilities of an individual by operationalizing them into test scores. An individuals' score on an assessment is evaluated in relation to a distribution of scores produced by others in either the same grade level, age cohort, or other shared group characteristic. The primary purpose of any assessment program is to take the information it gathers from test takers and use it to make evaluative judgements on both individual performance and group achievement. In order for an assessment program to adequately measure and evaluate what it is intentending to, there are two critical components of program design that must be incorporated during the preliminary stages of an assessment's development: 1) establishing performance level descriptors and 2) standard setting procedures to establish achievement level descriptors. When integrated effectively in an assessment design process, these two processes should inform one another on the necessary cut scores for the levels of achievement across a tests' score distribution. The cut score, or the minimum achievement score needed to meet a threshold, is developed in a standard-setting process that depends heavily on the definition for each level of performance.

Assessment programs can vary tremendously in their formats, purposes, and methodologies. The reason for this is because assessment program development is critical not only for the psychometric properties of the test items themselves, but also for the larger discussion on validation and implications of what tests are used to measure and what their scores mean to a larger social institution. The two types of assessment programs discussed in this paper are educational assessment programs, such as formative and summative assessments, as well as standardized testing programs such as achievement tests and aptitude tests. The chosen context of my topic in the latter half of this paper pertains to a field of research within the realm of educational psychology that I find quite intriguing: how to accurately identify a gifted student through measurement of mental knowledge – either through educational assessments or/and standardized assessments. Recent empirical evidence has shown widening excellence gaps in subgroups of students formally recognized as highly achieving, a phenomenon that I believe can be explained by an overall lack of special curriculum emphasis and/or teacher attention for students who are highly achieving in a traditional classroom setting, for these students are expected to not have inherent or unique academic and/or socio-behavioral struggles simply because they demonstrate high performance consistently. In brief, and at the risk of making a tacit statement, the importance of PLDs, ALDs, and cut scores for standard setting procedures becomes increasingly apparent when these processes may be what separates a child from the necessary educational services they need to be successful.

Performance Level Descriptors

Over the last two decades, performance level descriptors have amassed significant recognition by test developers and educational researchers alike as a result of the *No Child Left Behind Act* (2002). The purpose of the NCLB was for each state to formally adopt achievement descriptors for each grade and subject. In order to do so, the *NCLB's* statutes required that all assessment programs developed under the act to include at least three performance levels, and one of the levels was required to denote an average or standard benchmark per the subject's (i.e. math or reading/language arts) grade level requirements. The rationale behind a three-level requirement was to establish a level known as "proficient" via PLDs that were as well as at least one level both above and below this proficient category. Consequently, during the first decade of the 21st century, scores on state assessments (and any other standardized test developed under the *NCLB* act) were communicated as arbitrary category names rather than numeric values. As for the everyday educator, test taker, or parent, a common misconception was engendered regarding score feedback and how it translated to an individual's overall competency. The cut scores partitioning the levels of performance were – and still commonly are – misinterpreted as a descriptor of overall test achievement and subject aptitude, thus the actual descriptions of the levels themselves were quite literally lost in translation. To simply discern the difference between cut scores and PLDs, consider the former as quantitative and the latter as qualitative; to explicate the flow of their causal relationship, consider it one of exploratory nature with the qualitative informing how to develop the quantitative. The PLDs can be thought of as written checklists with explicitly defined items that must be met checked-off in order for a certain score value to be placed in a particular category.

In an article published by the *National Center for the Improvement of Educational Assessment* entitled "A Guide to Understanding and Developing Performance-Level Descriptors (2008)," author Marie Perie delineates the root-cause of discussions on why student scores at or above "proficient" on state math and/or reading assessments varied at such high degree across the United States during the early 2000's. According to Perie, the debates over why certain states, grade levels, age cohorts, genders, etc., had more students meeting "proficient" standards on state assessments in comparison to others stemmed from an inane point of contention over cut scores: "...most of these discussions center on the leniency or rigor of the cut score …[y]et, the cut score is developed in a standard setting process that depends heavily on the definition for each level of performance" (Perie 15). While PLDs are both necessary for and informative to the development of an assessment program (e.g. test items, cut-scores, reporting measures), when PLDs are inconsistent across several states' assessments, are poorly written (i.e. are ambiguous), or have too many levels (i.e. more than five), then PLDS were more likely the cause for such disparate scores denoting proficiency. To demonstrate the variations of how PLDs are written state-by-state, *Table 1* lays out the PLDs developed by state policymakers and education boards in Pennsylvania, Alabama, and Arizona. The key differences to note between the PLDs of these three states are how they vary in terms of each level description's title, length, word choice, and overall specificity.

Table 1.

Performance Level Descriptors (2008) for State Assessments to Determine Content Proficiency

PLD	Pennsylvania	Alabama	Arizona
Level 4	**Advanced:** Superior academic performance indicating an in-depth understanding and exemplary display of the skills included in Pennsylvania's academic standards	**Level IV:** Exceeds academic content standards	**Exceeds the Standard**: This level denotes demonstration of superior academic performance evidenced by achievement substantially beyond the goal of all students
Level 3	**Proficient:** Satisfactory academic performance indicating a solid understanding and adequate display of the skills included in Pennsylvania's academic standards	**Level III:** Meets academic content standards	**Meets the Standard**: This level denotes demonstration of solid academic performance on challenging subject matter reflected by the content standards. This includes knowledge of subject matter, application of such knowledge to real-world situations, and content-relevant analytical skills

			Attainment of this level is the expectation for all Arizona students
Level 2	**Basic:** Marginal academic performance, work approaching, but not yet reaching, satisfactory performance, indicating partial understanding and limited display of the skills included in Pennsylvania's academic standards	**Level II:** Partially meets academic content standards	**Approaches the Standard:** This level denotes understanding of the knowledge and application of the skills that are fundamental for proficiency in the standards
Level 1	**Below Basic:** Inadequate academic performance that indicates little understanding and minimal display of the skills included in Pennsylvania's academic standards	**Level I:** Does not meet academic content standards	**Falls Far Below Standard:** This level denotes sufficient evidence that the prerequisite knowledge and skills needed to approach the standard have not been met. Students who perform at this level have serious gaps in knowledge in skills related to Arizona's academic standard

Source: definitions from "A Guide to Understanding and Developing Performance-Level Descriptors" (Perie 2008), National Center for the Improvement of Educational Assessment. Page 17

A similarity seen in *Table 1* is between the word choice implemented by Pennsylvania and Arizona in Levels 3 and 4, particularly with the use of "superior" (level 4) and "solid" (level 3) by both states. Conversely, word choice is also a very apparent difference when comparing the titles for level 2 and level 1; Arizona's title and word choice shift seems much more drastic compared to Pennsylvania, especially because using terminology such as "falls far below" is subjective and ambiguous. Likewise, Arizona's phrasing in the definition for level 1, such as in the sentence "...serious gaps in knowledge…" is much more dramatic compared to the other two states.

Table 1 represents only three of the 29 states (as of 2008) that use a four-level system for their performance descriptions, however, according to the article published by Perie in 2008, approximately 10 states use the three-level system (Advanced, Proficient, Basic) recommended by the *NCLB* legislation, whereas the remaining 11 states use a five-level system. California is one example of a state that implements five performance levels via incorporating an additional level below proficient. In contrast, Delaware also uses five levels, however the state chose to incorporate an additional level above proficient. To demonstrate how cut scores can vary based on the number of PLDs for their assessment programs, consider two state assessments with verbatim PLDs for “advanced” and “proficient,” but one state’s assessment program has four PLDs because it divided level 1, or the level labeled “basic,” into two independent levels – “basic” and “below basic.” By adding a fourth level, the distribution of cut scores would be different on both assessments, as the test with four PLDs would have a narrower score range per level category.

The guidance prescribed by the National Assessment for Educational Progress (NAEP) for developing PLD’s is the approach that is the most valid and most supported by research, according to Perie. Developing PLD’s for establishing achievement levels entails “first specifying the numbers and names of the levels, next drafting policy definitions, and then fleshing out the policy definition with full descriptors for each subject and grade level … Ultimately, these descriptors communicate both the policy behind the meanings of labels such as “proficient” as well as the content expectations for each subject and grade assessed” (Perie 16). The second step of this approach is where standard setting procedures come in, as they are the processes by which performance and achievement descriptors are established thus fully developing the PLD into its full form.

Standard Setting Procedures

In Chapter 10 of *The Handbook of Test Development* (2006) entitled “Standard Setting” author Gregory Cizek delineates standard setting procedures, which are critical processes to the development of any assessment program for they entail setting performance standards per each performance level descriptor (PLD). Through operationalizing the PLDs via standard setting procedures, performance standards are set which, in effect, establish the assessment’s cut scores for measurement. There are several different procedural and statistical methodologies for standard setting; Cizek describes the procedures themselves as “...perhaps the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other” (Cizek 224). For the purpose of this overview and in relative context to the programs chosen for critical review and analysis in part two, the two standard setting procedures that will be discussed are traditional standard setting and Embedded Standard Setting.

Traditional standard setting is more procedural in nature. Cizek defines traditional standard setting as “the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance” (Cizek 226). This procedure of setting performance standards requires PLDs to be drafted and established first by policymakers; the chronological procedure that ensues is to first have a workshop facilitator guide a panel of experts (also referred to as panelists, judges, context experts) through the test content. A common approach used by standard setting panelists following traditional procedures is to use what is known as the Bookmark Method. Panelists first study test items in booklets ordered in difficulty, then their “primary job is to place a bookmark at the first location in the ordered item booklet such that students who demonstrate mastery of the content reflected by the items before the bookmark can be considered proficient” (Lewis 2020). In essence, traditional standard setting follows a due process procedure led by a panel of expert judges; this panel makes formative judgements on the difficulty of test items in relation to the pre-established PLDs: “… In most cases, the PLDs are defined by the panelists in order to describe the minimally acceptable test taker performance (Tannenbaum & Katz, 2013). It is this minimal acceptable performance, typically denoted using the Bookmark Method, that constitutes the cut score for proficient. As stated in the previous section, depending on the number of levels (i.e. PLDs), the panel will have to establish the corresponding number of cut scores.

The main issue that arises with traditional standard setting procedures is that establishing cut scores for more than two levels lowers the reliability scores of the panelists. Secondly, both Cizek (2006) and Lewis (2020) agree that traditional standard setting procedures can conflate their intended tasks and goals if PLDs are not established prior to setting performance standards and cut scores. Perhaps traditional standard setting procedures are only useful for “the conceptual nature of the endeavor” as described by Kane in Cizek’s chapter: “it is useful to draw a distinction between the passing score, defined as a point on the score scale, and the performance standard, defined as the minimally adequate level of performance for some purpose … the performance standard is the conceptual version of the desired level of competence, and the passing score is the optimal version” (Cizek 2006). Conceptual and/or optimal versions of an expected or desired performance level are not always realistic when developing assessment programs. It is more realistic that an educational assessment not be judged for proficiency by a panel of judges who are experts on standardized testing content and/or its methodologies, for this makes an assessment highly susceptible to bias and thus to having poor psychometric properties – especially low external validity. In Daniel Lewis's research paper, "Embedded Standard Setting: Aligning Standard Setting Methodology with Contemporary Assessment Design Principles," he argues that traditional standard setting using the bookmark method is both inherently biases and completely lacks any form of validity: "[test] items are rendered in a booklet according to some measure of empirical difficulty prior to being evaluated by panelists… [this form of empirical data] provides no guarantee that the cut scores or, more critically, the KSAs [knowledge, skills, aptitudes] measured by the items used to set them, reflect the descriptors intended to define the achievement levels established by the cuts" (Lewis 2020).

Embedded Standard Setting (ESS) procedures are part of the principled assessment design framework; over the last two decades, recent advances in test design under this framework have rendered traditional standard setting procedures antiquated and barren in comparison. Lewis notes two principled assessment design frameworks – Evidence Centered Design and Assessment Engineering – that have influenced the methodologies for ESS procedures, which Lewis argues to be more transparent, thoughtful, and rigorous than traditional procedures. Whereas traditional standard setting utilizes either the Bookmark or modified Angoff method, ESS procedures are different in how they allow achievement level descriptors (same concept as PLDs) to be modified and adjusted throughout the item-level rating process. Through the collection of empirical data through field and pilot studies of actual test items, either through participant observation, interviews, questionnaire feedback, etc., data on examinee performance can be used in a confirmatory process to either confirm or challenge decisions made by panels of item-writers dictating which items belong in which performance level. What makes ESS procedures unique and separate from traditional standard setting is that it uses student-specific knowledge, skills, and aptitudes to make more evaluative judgements on realistic student achievement standards.

Integrating PLDs and Standard Setting Procedures

In assessment program design, integrating PLDs and standard setting procedures into the process requires that each achievement level be clearly defined and its expected standards communicated, agreed upon, and confirmed by empirical evidence. As stated previously, the *NCLB* (2002-2015) required assessment program developers to implement performance levels by emphasizing a level marked as proficient, thus requiring that at least one level exist both above and below in order to provide scaling context for cut score distributions. While the *NCLB* act is no longer in effect, assessment programs still continue to derive meaning from scores. As stated by Lewis in his critical analysis on ESS procedures, "it is this association between [test] item, ALD, and achievement level that imbues the test score with meaning and allows for valid interpretation of said score" (Lewis 2020).

The concept of validity in relation to assessment programs used to measure mental knowledge and make evaluative judgements based on the operationalizations of these measures is a long-standing debate in the field of education and psychometrics. Going back over a century, Truman Lee Kelley, known as the champion of the myth of mental measurement, noted that previous expressed agonies over the measurement units produced by testing were unnecessary debates: "starting with units however defined, if we can establish important relationships between phenomena measured in these units, we have proceeded scientifically. The choice of unit is purely a question of utility … While the fact that test scores relate to important criteria needs explaining, the assumption that they do this only because they measure something presumes that the relevant attributes are quantitative" (Michell 225). Essentially, test developers for assessment adopted standard setting procedures such as PLDs, ALDs, and cut scores, in order to make more sense of how mental knowledge is operationalized via testing. And while a singular test score should never be the sole determinant of an individual's proficiency in any field of knowledge, skill, or ability, humans use testing to make somewhat informed decisions – because decision-making is not only human nature but also inevitable in most cases. As stated by Cizek on the topic of test scores and their practical and real-world implications, "Exceptional performance on a high school graduation test is of no avail if the student has not accumulated the requisite credit hours, GPA, and met other requirements. Although clearly not the sole criterion, it is certain that information yielded by tests plays an important part in the decisions as diverse as placement in a remedial or gifted program …" (Cizek 228).

PLDs and Standard Setting in Context of Gifted Student Identification and Measurement

Prior to the turn of the 21st century, it was the responsibility of either state education systems, local school districts, or grassroots organizations to spearhead and mobilize efforts that would combat the lack of federal support for gifted and talented (GAT) educational resources nationwide. Resources such as the necessary research for policy change, adequate assessment program design frameworks to identify and measure giftedness, and monetary funds for program development within public schools became even more scarce following the *NCLB* act of 2002. Nowhere in its legislation were the learning goals, educational standards, or program recommendations for this academically advanced group ever mentioned. An overemphasis on identifying those students who were either at or below "proficient" per each states' achievement levels had an inverse and detrimental effect on students scoring on the polar end of the achievement spectrum. For example, a graphic portrayal of gifted students over a span of 10 years demonstrated how the *NCLB* act caused GAT students to languish in U.S. schools. The graphic displayed data from 1998 to 2008, with its curve denoting a "10-year flat profile of growth for high-ability learners, while lower ability students showed significant gains" (VanTassel-Baska 2018).

An approximate 5%–7% of all K-12 students in each state are estimated to meet state achievement measures that would categorize them as gifted, however very few states have policies in place to meet the educational needs of this group. Correspondingly, it was estimated that. As of 2022, there are over six million GAT students in the U.S. who are not receiving adequate educational services. In educational institutions where gifted services are not available, underachievement among the subgroup is rampant, and excellence gaps are widening between socioeconomic and demographic groups. Even with the Every Student Succeeds Act (ESSA) in place as of 2015, there is still a general lack of concern at the federal and state levels to meet the needs of the academically advanced. Empirical evidence from localized studies has found that gifted students are also at a "greater risk for dropping out of high school or underachieving if their needs are not met, with 20% of high school dropouts identified as gifted and more than 30% as underachieving" (VanTassel-Baska 2018). Underachievement has a major point of contention for educational researchers, teachers, and parents, as GAT students are not the "typical" group normally considered to be at risk for academic underachievement. This misconception is what has led to widening excellence gaps for the highly achieving; more specifically, these excellence gaps appear to be worsening between students of differing demographic and socioeconomic groups. A policy document issued in 2012 noted research on the plight of poor gifted student education across the U.S., stating that achievement gaps are substantial for gifted students whose families are of lower socioeconomic (SES) status. Correspondingly, a more recent analysis (Plucker et al., 2013) comparing the academic achievement trajectories between gifted students of high SES and low SES found that achievement gaps exist in all academic subjects and widen each year beginning in elementary school and continue to do so throughout highschool.

Over the last half century, the U.S. education system has taken on a laissez-faire approach to its governance structure for gifted education, placing the educational needs of GAT students in the hands of state legislatures and state departments of education. Every facet of gifted education policy – from determining its PLDs in state assessments to the allocation of program funds to districts state-wide – is left entirely up to policymakers at the state level. Because of the fact that there is still no definition, mandate, or funding process in effect at the federal level, state's are not required to enact policy on the topic or even address gifted education policy at all: "Gifted education policy at the state level is tied to the rules, statutes, codes, and regulations adopted by state legislatures, interpreted by state school boards of education and state departments of education, and implemented by local school districts" (VanTassel Baska 2018).

In the context of how gifted education policy is interpreted, implemented, and mandated through state policy, the first step for any state education board is to design an assessment program (or programs) that adequately identifies gifted students through reliable and valid measures of achievement, ability, and performance. Therefore, the use of performance level descriptors and standard setting procedures are critical components to the design process of assessment programs when utilized in the context of addressing the educational needs of GAT student populations.

Part Two: Program Critique

Assessment Program: Project STAR

Project STAR was a policy initiative implemented by the state of South Carolina in 2000. The project was designed to include more low income and minority students in gifted programs by lowering the cut scores (i.e. widening the acceptable score threshold) that corresponded with the state's PLDs and performance standards denoting giftedness on performance-based assessments in both math and reading. The initiative developed its own gifted identification protocol that combined existing identification criteria – such as previous state assessment scores – in addition to performance-based assessments that "allowed students to be considered for gifted programs if they met a lower minimum score level on an ability or achievement test: 94th [percentile] instead of 96th [percentile] on achievement; 90th [percentile] instead of 96th [percentile] on ability" (Van-Tassel Baska, 2018).

Students were selected based on a domain-specific strength in one or more areas of the assessment by utilizing previous identification protocol and new performance-based assessment tasks. The newly identified students were placed in gifted programs alongside peers identified using traditional PLDs and testing instruments (i.e. comparison group). The sample group's achievement scores and ability scores on performance-based assessments were tracked over two years and compared to their peers and to state achievement measures. During the first field test of the instrument in 2001, which included a sample participant pool of more than 4,000 K-8 students across multiple public school districts in the state, lowering cut scores for achievement and ability by two percent and six percent, respectively, resulted in identifying a group of gifted learners who were "12% of low SES and 14% African American" (Van-Tassel Baska & Johnson, 2002). Several more field tests were conducted in order to develop and revise the performance-based assessment tasks, and student achievement was tracked and recorded over a two year period: "Performance data suggested that these students were performing at proficient and above levels on state achievement measures in their area of strength 2 years in the program…this finding was comparable with more advantaged students identified under the existing identification system after 1 year in the program" (Van-Tassel Baska, 2018).

According to a report published by the same researchers more than 15 years after the initial study, the results would still suggest that “lowering the threshold cutoffs on traditional instruments, coupled with employing domain-specific performance-based measures, results in larger numbers of low income and minority students being identified and served in gifted programs with no negative impacts on achievement nor on gifted program practices…” (Van-Tassel Baska, 2018). Project STAR is therefore a clear demonstration of how slight alterations to PLDs and performance standards can entirely change the distribution of students in each performance level state-wide. By simply lowering the cut score associated with gifted performance standards defined by the state’s PLDs, further longitudinal analysis showed that “a range of 10% to 14% more low-income and African American students became eligible for gifted programs state-wide” (Van-Tassel Baska, 2018).

The focus of Project Star was based on “the need to identify more African American children as gifted and talented to meet an order from the Office for Civil Rights” (Van-Tassel Baska, 2018). The rationale for developing performance assessment tasks specifically to augment the identification of more economically disadvantaged and minority students for gifted programs in one state is certainly controversial for numerous reasons. However, in a more pragmatic lens, opening doors to educational opportunities should always be evaluated and critiqued lightly, yet using the same processes in a different socio-cultural context or with a different population should be evaluated critically and heavily before doing so. One positive methodological aspect of Project STAR that proved to be beneficial from a test developer’s perspective is that it implemented ESS methodologies in order to generate ALDs that were more prescriptive of what students at each achievement level should know and be able to do. In the same vein, these ALDs would look different for different student populations and thus verbatim replication of Project STAR’s methodologies would not be advised.

Assessment Program #2: A Critique of MCAS and the Implications of Implementing gifted PLDs and performance standards

A critical analysis of the Massachusetts Comprehensive Assessment System, or MCAS for short, is especially relevant to the topic of PLDs, ALDs, and cut scores when discussed in the context of using assessment programs to identify and measure giftedness because MA’s complete lack of state policy towards meeting the educational needs of this student subpopulation. The policy practices currently in effect are what typify MA as an outlier in comparison to other U.S. states. Nearly every other state in the country defines giftedness, however MA is one of the few that does not provide any definition at all (see appendix: Table 2). Likewise, according to the *State of the States in Gifted Education* (Ansel 2019), MA has no explicit mandates in place to either identify or serve GAT students state-wide, whereas in 32 states these mandates do exist in state educational policy legislatures. Further validating the argument that MA is an outlier in its approach to gifted policy, when comparing MA's gifted policy to the state policies of its 15 closest economic competitor (Appendix Table 3), it is increasingly apparent that there are economic resources available to mobilize efforts at the state level in some capacity.

While it should be noted that MCAS assessment programs do not currently serve as tests with an evaluative purpose to identify and measure giftedness, I chose to critically analyze the current standard setting procedures for MCAS's PLDs, ALDs, and cut scores as they are delineated in the 2019 report, "Standard Setting Meeting Executive Summary for MCAS Grade 10 ELA, Grade 10 Math, and Grades 5 and 8 STE" (https://www.doe.mass.edu/mcas/tech/) in comparison to a report by Dr. Dana Ansel published in 2019 entitled "Gifted Education in Massachusetts: A Policy and Practice Review." This report was commissioned by The Department of Elementary and Secondary Education the year prior and was presented to the Massachusetts State Legislature with the purpose of both informing policy makers on the ramifications of student outcomes and overall well-being when gifted programs are not available (especially at the elementary level), as well as to provide recommendations on how the state should proceed for adopting gifted policy.

The current environment within MA gifted educational programs are few and far-between options for academically advanced students; unless a student is of high SES (higher tax rates in high SES neighborhoods result in wealthier school districts with expendable resources for supplemental programs) gifted educational services are essentially nonexistent. It won't be until a gifted individual reaches the middle or even high school level where differentiation from a normal classroom setting is possible (i.e. advanced placement classes or opportunities to take elective courses of interest), even if that student has, for instance, consistently demonstrated advanced scores on MCAS reading assessments that should warrant higher grade level education through content acceleration. To conceptualize the realities of a needs assessment for GAT educational programs, the report by Ansel estimated that, based on a comparative value analysis of states with similar student population compositions, between six to eight percent of all K-12 students in MA are estimated to be truly "gifted" by traditional standards but are not yet not identified as such via MA state policy. To conceptualize this percentage, six percent translates to roughly 57,000 total students in the state of MA, those of which are not receiving adequate programs and/or services to meet their educational needs.

Ansel's report is not only a comprehensive overview but also a critical analysis explaining the need for gifted education policies and practices and their implementation at the state level. The common theme in the report is that gifted programs would allow MA public education systems to address growing excellence gaps prevailing between demographic and socioeconomic groups in hopes of mitigating these disparities. It should be noted that "excellence gaps" and "achievement gaps" are not synonymous in this context, as achievement gaps could easily be confused with not meeting certain cut scores prescribed by the ALDs and PLDs. Excellence gaps refer to differences between subgroups of students performing at the highest levels of achievement, and these are what have garnered attention more recently at the national level: "...researchers find that very few low-income students score at the advanced level on any national tests. Similarly, they document large excellence gaps between students of different races and ethnicities. Massachusetts has some of the largest excellence gaps in the country, despite the fact that the percentage of students in Massachusetts scoring advanced on state and national assessments has increased" (Plucker & Peters, 2016). This finding is one that should be addressed

Both a challenge and a limitation of Ansel's report is the availability of valid and reliable data for making formative and summative judgements on a state-wide gifted policy. In particular, the absence of a definition of giftedness makes determining the proportion of academically advanced students that are actually "gifted" virtually impossible. To overcome obstacles in information availability, Ansel notes that she first focused on qualitative data by reviewing state policies as well as collecting information from parents of the academically advanced via semi-structured interviews. To combat the fact that MA does not define or collect data on giftedness in students, Ansel instead focused on “analyzing the academic trajectory and social-emotional well-being of academically advanced students based on their math MCAS scores … [all] of this information is valuable in painting a picture of gifted education in Massachusetts, but it is nonetheless limited (Ansel 2019). Through evaluative assessments of the few districts in MA that do have gifted programs, noting that gifted programs exist in the school districts where families are of a higher SES; this is due to The author notes that it is imperative that MA take action at the state level in order to meet the educational needs of GAT students in poorer districts across the state. In correspondence with PLDs and standard setting procedures, below are the seven steps recommended by Ansel that state educators and policymakers follow in order to mitigate excellence gaps between demographic and SES groups within districts:

> “The central message of this report is that the current hands-off approach of Massachusetts, with few gifted programs and not much attention to gifted education, is not serving advanced and gifted students well. In particular, when we tracked one statewide cohort of academically advanced students, we found stark differences in the academic outcomes of Black, Hispanic, and/or low-income students, as compared with white and Asian students. Our analysis documented the widening of the excellence gap between 3rd and 6th grade
> The research findings from this report lead to the following recommendations:
>
> **Create a statewide taskforce, which will;**
>
> 1. **Define giftedness and measures to assess giftedness;**
> 2. Determine most effective way to collect data on gifted students;
> 3. **Consider best practices of other states and districts;**
> 4. Establish state policy and guidelines on acceleration;
> 5. Track and report on the excellence gap; identify and implement strategies to close it.
> 6. Include instruction on the learning needs of gifted students as part of teacher training for all teachers; and
> 7. Hire staff at the Department of Elementary and Secondary Education with expertise in gifted students and gifted education.”
>
> **Source: Ansel 2019**

In order to adequately critique the MCAS program in the context of gifted achievement standards and performance descriptors, the test’s design and development procedures will be briefly discussed in order to fully understand the ramifications and practical implications of implementing PLDs, performance standards or ALDs, and cut scores for measuring and identifying the GAT student population in MA. A simplified model (See Appendix: *Table 4*) found on MCAS website (https://www.doe.mass.edu), provides a visual representation of the assessment's design and development procedures for generating test items. In comparison to both the article on embedded standard setting procedures (Lewis 2020) referred to previously and to the performance task development methodologies utilized in the STAR program, the standard setting procedures implemented by MCAS panelists and board members appears to be thorough and transparent in terms of its continual evaluation of test items' reliability and validity in relation to ALDs. MCAS developers use the yes/no modified Angoff method, which was mentioned in Lewis's article as being a statistical method for ESS that is more valid and reliable that traditional standard setting procedures. Likewise, the report, "Standard Setting Meeting Executive Summary for MCAS Grade 10 ELA, Grade 10 Math, and Grades 5 and 8 STE" (https://www.doe.mass.edu/mcas/tech/), pays particular attention to biases that could arise during standard setting workshops and/or by team members themselves as individual evaluators, such as observer bias and confirmation bias, and the report delineates the appropriate steps that were taken in order to account for errors that could have been from biased interpretations of test items or summative judgements on these items.

What the MCAS could implement moving forward in order to work alongside assessment developers for identifying and measuring gifted K-12 students in Massachusetts are two practices that pertain to the drafting and establishing of PLDs prior to standard setting. As noted in the PLD section in Part One, some states have five PLDs for assessments. One state in particular that stood out for fitting the criteria for being similar to MA was Delaware's PLD structure. DE has five levels (as compared to MA's four), but instead of adding an additional level below the "proficient" category, which is the more common method, DE added one level above proficient. Because of the fact that proficient test scores naturally cause a normal score distribution (i.e. CLT bell curve), splitting the fourth quartile into two categories could be a way to separate the academically advanced students from the gifted students. This would also align with other standardized measures of giftedness, as they usually have a benchmark between 90th percentile to 95th percentile (and above). Consider the bar for ELA Grade 10 in *Table 5 (*see appendix); if an additional category was adopted as a fifth PLD above the proficient level, then the 14% of students would be placed in either level based on their ALDs and corresponding cut scores. This could be one way to use PLDs as the foundation for establishing gifted identifiers in assessment programs across MA.

The second process for identifying gifted students using PLDs is to follow a similar one such as was used in Project STAR in South Carolina. To reiterate a finding mentioned above, “MA has some of the largest excellence gaps in the country, despite the fact that student percentages scoring advanced on state and national assessments has increased” (Ansel 2019). This could mean that lowering the cut score threshold and altering corresponding PLDs and ALDs for certain school districts of lower SES or of particular demographic groups could be advantageous if done at a district-by-district level. The reason for a more localized level of structure change is to account for that particular student populations’ knowledge, skills, and abilities relative to state-level PLDs. Lowering the threshold for cut scores denoting giftedness could therefore be a possible way to mitigate widening excellence gaps between subgroups in the population who are scoring as proficient versus advanced.

References

Ansel, D. (2019, August 20). Gifted education in Massachusetts: A policy and practice review. https://archives.lib.state.ma.us/handle/2452/807459

Downing, S. M., Haladyna, T. M., & Cizek, G. (2006). Standard Setting. In *Handbook of Test Development* (pp. 225–260). essay, Taylor & Francis.

Lewis, D., & Cook, R. (2020). Embedded standard setting: Aligning standard‐setting methodology with Contemporary Assessment Design principles. *Educational Measurement: Issues and Practice*, *39*(1), 8–21. https://doi.org/10.1111/emip.12318

Massachusetts Department of Elementary and Secondary Education. (n.d.). *MCAS technical documents*. MCAS Technical Documents - Massachusetts Comprehensive Assessment System. https://www.doe.mass.edu/mcas/tech/default.html

Mitchell, J. (2009). Invalidity in Validity. *The Concept of Validity: Revisions, New Directions, and Applications*, 72–91. https://doi.org/10.4324/9780367854324-7

Perie, M. (2008). A guide to understanding and developing performance‐level descriptors. *Educational Measurement: Issues and Practice*, *27*(4), 15–29. https://doi.org/10.1111/j.1745-3992.2008.00135.x

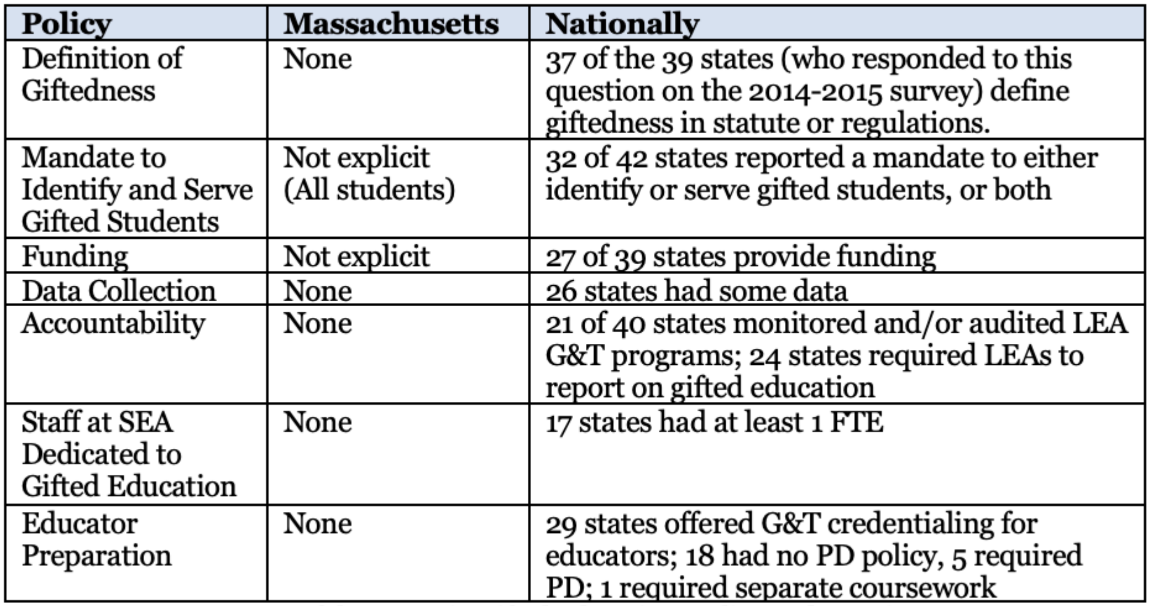
VanTassel-Baska, J. (2018). American policy in Gifted Education. *Gifted Child Today*, *41*(2), 98–103. https://doi.org/10.1177/1076217517753020

VanTassel-Baska, J., Johnson, D., & Avery, L. D. (2002). Using performance tasks in the identification of economically disadvantaged and minority gifted learners: Findings from Project Star. *Gifted Child Quarterly*, *46*(2), 110–123. https://doi.org/10.1177/001698620204600204

Appendix

Table 2.

MA Gifted Policies Compared to Other State Policies

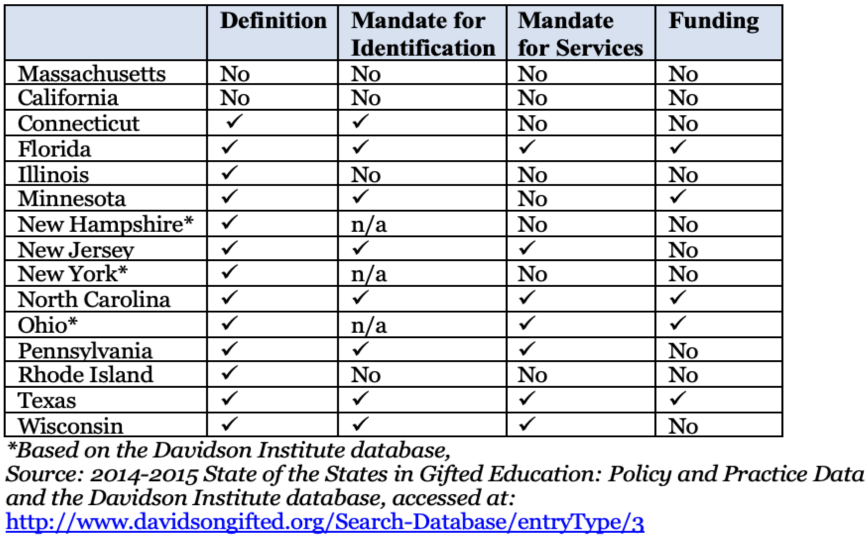


Policy	Massachusetts	Nationally
Definition of Giftedness	None	37 of the 39 states (who responded to this question on the 2014-2015 survey) define giftedness in statute or regulations.
Mandate to Identify and Serve Gifted Students	Not explicit (All students)	32 of 42 states reported a mandate to either identify or serve gifted students, or both
Funding	Not explicit	27 of 39 states provide funding
Data Collection	None	26 states had some data
Accountability	None	21 of 40 states monitored and/or audited LEA G&T programs; 24 states required LEAs to report on gifted education
Staff at SEA Dedicated to Gifted Education	None	17 states had at least 1 FTE
Educator Preparation	None	29 states offered G&T credentialing for educators; 18 had no PD policy, 5 required PD; 1 required separate coursework

Source: 2014-2015 State of the States in Gifted Education: Policy and Practice Data

Table 3.

Economic Competitor States' Policies Towards Gifted Education



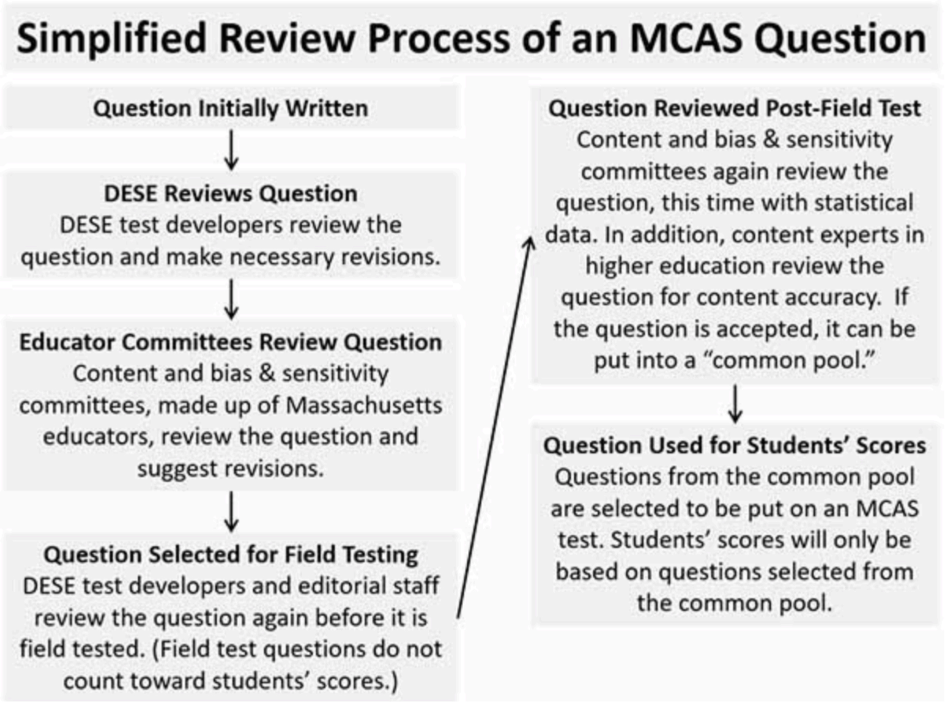
	Definition	Mandate for Identification	Mandate for Services	Funding
Massachusetts	No	No	No	No
California	No	No	No	No
Connecticut	✓	✓	No	No
Florida	✓	✓	✓	✓
Illinois	✓	No	No	No
Minnesota	✓	✓	No	✓
New Hampshire*	✓	n/a	No	No
New Jersey	✓	✓	✓	No
New York*	✓	n/a	No	No
North Carolina	✓	✓	✓	✓
Ohio*	✓	n/a	✓	✓
Pennsylvania	✓	✓	✓	No
Rhode Island	✓	No	No	No
Texas	✓	✓	✓	✓
Wisconsin	✓	✓	✓	No

**Based on the Davidson Institute database,*
Source: 2014-2015 State of the States in Gifted Education: Policy and Practice Data and the Davidson Institute database, accessed at:
http://www.davidsongifted.org/Search-Database/entryType/3

Massachusetts is an outlier in its approach to gifted students and gifted education. It is one of the few states in the country that does not have a definition for giftedness. It neither collects data on gifted students, nor is there a mandate to identify or serve gifted students. Other New England states are also outliers in their approach to gifted education, although every other New England state defines giftedness. Compared with its economic competitor states, Massachusetts and California are similar in their lack of definition or mandates for identification and services. The approaches of the other 13 states differ, with Florida, North Carolina, Ohio, and Texas providing funding in addition to mandates for identification and services.

Table 4.

Simplified Model of MCAS Test Item Development Design Process



Simplified Review Process of an MCAS Question

Question Initially Written

DESE Reviews Question
DESE test developers review the question and make necessary revisions.

Educator Committees Review Question
Content and bias & sensitivity committees, made up of Massachusetts educators, review the question and suggest revisions.

Question Selected for Field Testing
DESE test developers and editorial staff review the question again before it is field tested. (Field test questions do not count toward students' scores.)

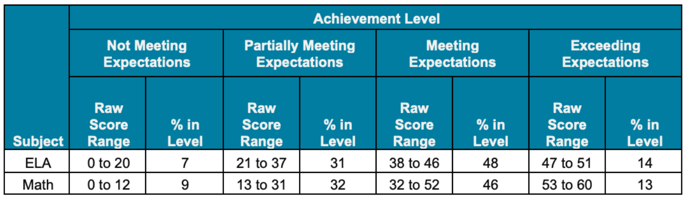
Question Reviewed Post-Field Test
Content and bias & sensitivity committees again review the question, this time with statistical data. In addition, content experts in higher education review the question for content accuracy. If the question is accepted, it can be put into a "common pool."

Question Used for Students' Scores
Questions from the common pool are selected to be put on an MCAS test. Students' scores will only be based on questions selected from the common pool.

Source: https://www.doe.mass.edu/mcas/tech/default.html

Table 5.

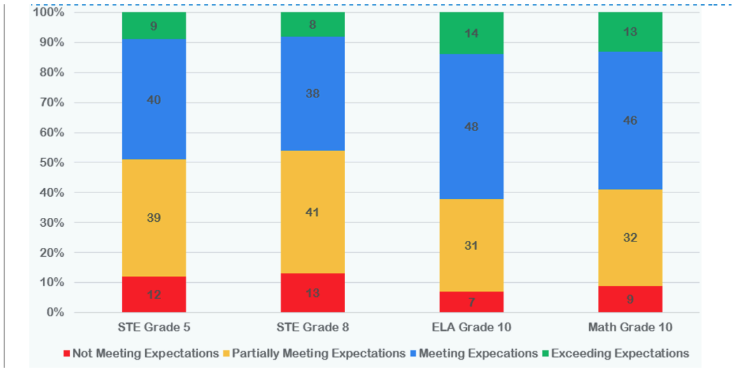
Impact Data Table for Grade 10 ELA Assessments and Graphical Display of Score Distributions for Four Level PLDs



	Achievement Level							
	Not Meeting Expectations		Partially Meeting Expectations		Meeting Expectations		Exceeding Expectations	
Subject	Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level
ELA	0 to 20	7	21 to 37	31	38 to 46	48	47 to 51	14
Math	0 to 12	9	13 to 31	32	32 to 52	46	53 to 60	13

Figure 3 presents the impact data from the final recommendations as stacked bar graphs.

Figure 3. Impact Data for STE, ELA, and Math Tests based on Final Recommendations



Source: https://www.doe.mass.edu/mcas/tech/default.html