Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1476

The Challenge of Capturing Higher-Order Thinking Skills at Scale

John Wang
University of Virginia

Beth E. Schueler
Stanford University

Higher-order thinking skills are important for K-12 students' long-term success. However, the lack of widely administered assessments designed to capture this construct has made it difficult to measure higher-order skills at scale. This paper examines the measurement properties of an approach to capturing higher-order skills using extant statewide standardized testing data. We use item-level data from Massachusetts's English Language Arts exams for over two decades (2001 to 2023) and pair this with a novel coding which flags whether an item captures higher-order skills based on the Webb's Depth of Knowledge framework. Overall, we find that this is a challenging approach for separating out higher-order from lower-order skills because state assessments were designed to measure unidimensional constructs. However, there are some cases where the data do appear to distinguish higher-order from lower-order skills in testing years prior to the introduction of Common Core-aligned assessments. Data from these grade-years could be used to isolate and study higher-order skills. However, we encourage researchers to proceed with caution when relying on this approach and to also pursue alternative avenues to measuring these important skills.

VERSION: May 2026

Suggested citation: Wang, John, and Beth E. Schueler. (2026). The Challenge of Capturing Higher-Order Thinking Skills at Scale. (EdWorkingPaper: 26-1476). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/asn1-ev31

The Challenge of Capturing Higher-Order Thinking Skills at Scale

John Wang, University of Virginia
Beth E. Schueler, Stanford University

Abstract: Higher-order thinking skills are important for K-12 students' long-term success. However, the lack of widely administered assessments designed to capture this construct has made it difficult to measure higher-order skills at scale. This paper examines the measurement properties of an approach to capturing higher-order skills using extant statewide standardized testing data. We use item-level data from Massachusetts's English Language Arts exams for over two decades (2001 to 2023) and pair this with a novel coding which flags whether an item captures higher-order skills based on the Webb's Depth of Knowledge framework. Overall, we find that this is a challenging approach for separating out higher-order from lower-order skills because state assessments were designed to measure unidimensional constructs. However, there are some cases where the data do appear to distinguish higher-order from lower-order skills in testing years prior to the introduction of Common Core-aligned assessments. Data from these grade-years could be used to isolate and study higher-order skills. However, we encourage researchers to proceed with caution when relying on this approach and to also pursue alternative avenues to measuring these important skills.

Keywords: critical thinking, higher-order thinking, analytical thinking, educational testing, educational measurement, educational assessment

Acknowledgements: This work has been possible due to funding from the Walton, Gates, and Joyce Foundations through the Student Upward Mobility Initiative at the Urban Institute. We were also generously supported by the U.S. Department of Education's Institute of Education Sciences through Grant No. R305B200005 to the University of Virginia (UVA) as well as the Bankard Fund for Political Economy and Education and Human Development Dean's Research and Development Fund at UVA. We thank James Soland, Daphna Bassok, Allison Atteberry, Daniel Player, participants in the UVA EdPolicyWorks Seminar Series, and the Annenberg EdWorkingPaper Review Board for valuable feedback on this project. UVA students Jenna Phillips, Isabelle Saillard and Tess Lenihan provided excellent research assistance. Finally, we thank the Massachusetts Department of Elementary and Secondary Education for making public much of the data we have analyzed here.

Introduction

Higher-order thinking skills ("HOTS")—meaning analytical, reasoning, and problem-solving abilities—are important for K-12 students' long-term success and are likely to become even more valuable as the use of artificial intelligence becomes more ubiquitous. Despite the value of these skills, researchers have struggled to study HOTS at scale for broad, representative populations of students over extended periods of time. Prior efforts to capture HOTS have faced critical limitations, including a lack of widescale assessments that include a sufficient number of items measuring HOTS, unrepresentative samples, and brief years of test administration. One underexplored approach involves using statewide standardized exams administered at scale, directly coding individual assessment items based on the extent to which they assess higher-order thinking, and pairing these codes with student item-level performance data to create a measure of higher-order thinking skills.

This paper investigates the psychometric validity of such an approach. In particular, we use test item-level data from the statewide assessment used for accountability purposes in Massachusetts from over two decades (2001 to 2023) across two testing regimes—before and after the exams were aligned with the Common Core state standards. We generate a novel coding of all publicly available English-Language Arts (ELA) standardized assessment items that flags whether each item captures HOTS and pair the coding with student item-level performance information. We then test whether supporting validity evidence exists for using the measure to capture HOTS and differentiate them from lower-order thinking skills (LOTS). More specifically, we compare a series of measurement models designed to test different assumptions about the items coded as capturing HOTS. After identifying whether the use of models distinguishing HOTS from LOTS can be supported, we examine the models' characteristics and explore how well HOTS scores predict educational attainment outcomes as a form of predictive validity evidence.

The results indicate that using pre-existing statewide data to capture higher-order thinking is challenging, though not impossible depending on the test regime. Substantial evidence supports the use of a unidimensional measurement model in both pre- and post-Common Core-aligned testing regimes. This makes sense given these tests were designed to primarily capture a unidimensional construct of ELA ability. More cognitively demanding assessments, such as the Common Core-aligned "MCAS 2.0" assessments in Massachusetts include a larger share of items capturing HOTS and therefore likely—as a whole—better reflect HOTS than earlier assessments with a lower share of HOTS items. However, these MCAS 2.0 assessments do not provide a good source of information for distinguishing HOTS from LOTS because measurement models suggest the assessments are unidimensional. Even the pre-Common Core-aligned "MCAS 1.0" assessments are often not a good source of information for distinguishing HOTS from LOTS as they also are often unidimensional. However, some grade-year results do allow for differentiation between HOTS and LOTS when using the pre-Common Core assessments, suggesting these data can be used to study and compare HOTS and LOTS. Our work therefore shows both the possibility of using administrative data to distinguish higher-order from lower-order skills at scale along with key challenges and limitations researchers should consider before moving forward with such an approach.

What Are Higher-Order Thinking Skills?

We define "higher-order thinking skills (HOTS)," sometimes called "critical thinking skills," as any form of systematic thinking involving an individual's ability to make a claim or

take a task, identify evidence or data that are relevant to the claim/task, and provide an evaluation, judgment, or solution. While consensus on the precise definition has remained elusive in the field (Alexander, 2023), this definition synthesizes key features across a wide array of frameworks that have been used for thinking about these types of skills (e.g., Bloom et al., 1956; Facione, 1990; Hess et al., 2009). Some scholars refer to "higher-order skills" as an umbrella term covering 21st-century skills such as non-cognitive, social-emotional, teamwork, and abilities (e.g., Deming, 2022). Although these skills are undoubtedly important, the intra- or inter-personal components (e.g., grit and social skills) operate somewhat distinctively from the thinking skills (Koenig, 2011), leading to our focus here on the analytical, reasoning, and problem-solving aspects.

Scholars further debate the degree to which higher-order skills represent general or domain-specific skills. Although one extreme conceptualizes higher-order thinking as a generalizable skill applicable to any domain like history or science, researchers rarely take such a position. Instead, scholars tend to fall somewhere on a continuum between viewing HOTS as general skills that require prior foundational domain knowledge to fully demonstrate higher-order thinking (e.g., Facione, 1990) and conceptualizing each domain's higher-order thinking as entirely separate from other domains, for example, distinct history higher-order skills from science higher-order skills (e.g., Willingham, 2008). These theoretical debates have implications for the actual measurement of higher-order thinking, shaping whether the construct is represented as a generalizable skill or tied to the foundational domain knowledge.

Why Study Higher-Order Thinking Skills In K-12 Schools?

Higher-order thinking skills appear increasingly important for economic and civic success in adulthood. Descriptive research suggests that these competencies are in high demand in the labor market as compared to more routine cognitive tasks ("lower-order thinking skills", or LOTS) or manual skills (Autor, Levy & Murnane, 2003; Deming, 2017; Deming 2022), and are associated with a wage premium (Autor & Handel, 2013; Deming & Kahn, 2018) that has increased over time (Liu & Grusky, 2013). Furthermore, scholars point to the development of HOTS, particularly as it relates to civic reasoning, as uniquely important for ensuring equitable participation in the political process (Gutmann, 1999; Stitzlein, 2021). HOTS could also influence voter turnout simply by promoting educational attainment, one of the most consistent predictors of civic participation (e.g., Campbell, 2006; Gilens & Page, 2014).

K-12 schooling presents a key opportunity to shape individuals' HOTS. Individuals develop HOTS through explicit instruction of thinking principles, immersion in thought-provoking content, or a mixture of both (Abrami et al., 2008; 2015). While this developmental process may happen outside of school contexts, K-12 schools provide many structured opportunities to support higher-order thinking development through regular and advanced coursework, extracurricular activities, and more (e.g., Schueler & Larned, 2023). Therefore, it is important to measure HOTS during the K-12 years, when schools still have the opportunity to shape these skills. Unfortunately, it has been challenging for researchers to generate knowledge about K-12 students' HOTS at scale.

The Challenges of Examining Higher-Order Thinking Skills at Scale

Generating knowledge about HOTS—such as identifying average levels of HOTS among K-12 students, differences in HOTS between groups, the equitable or inequitable distribution of HOTS, and how educational systems could best promote HOTS—requires a K-12 HOTS

measure systematically administered across subgroups to wide populations of interest. However, assessments designed to capture HOTS have largely not been administered to broad, representative samples of K-12 students. Exams that are administered widely (e.g., statewide or nationally) do not typically capture HOTS. Scholars recommend that for a whole assessment's scores to accurately capture the HOTS construct, 50% or more of the items should measure HOTS (Darling-Hammond et al., 2013). Nationally representative exams—such as the National Assessment of Educational Progress (NAEP), Programme for International Student Assessment (PISA), Progress in International Reading Literacy Study (PIRLS), or Trends in International Mathematics and Science Study (TIMSS)—do not meet the 50% HOTS items threshold guideline (Yuan and Le, 2014; Herman et al., 2016), so scores generated from the overall assessment would not sufficiently capture HOTS. Other policy-relevant exams such as the Preliminary SAT (Esdale, 2021), SAT, ACT (Christopherson and Webb, 2018), International Baccalaureate (IB, Yuan and Le, 2014), and Measures of Academic Progress (MAP, Gareis et al., 2021) similarly include only a minor fraction of items measuring HOTS and are not always available for representative samples. Prior scholarship finds that the more widely administered state assessments used for accountability purposes in the U.S. also do not consistently include 50% of items capturing HOTS (Yuan and Le, 2012; Doorey and Polikoff, 2016). One exception that does meet the recommended threshold for share of HOTS items are the ELA Advanced Placement (AP) Exams. However, AP exams are administered to an unrepresentative population of disproportionately advantaged students who attend schools with access to AP coursework and exams (Iatarola et al., 2011) and who self-select to participate. Other assessments likely meeting the 50% criterion (see Liu et al., 2014 for a review), have been predominantly administered in isolated, postsecondary educational contexts.

To be clear, though assessments designed to capture HOTS have historically largely not been administered to broad, representative samples of K-12 students, Common Core State Standard (CCSS) aligned statewide assessments offer one effort to capture higher-order thinking on a wider scale. CCSS, and its aftermath, represented a shift in the rigor of state academic standards (Peterson et al., 2016). Unlike the previous No Child Left Behind-era assessments, which had very few items capturing HOTS (Yuan and Le, 2012), contemporary, CCSS-aligned statewide assessments offer a substantial increase in the cognitive demand relative to earlier pre-Common Core testing regimes (Appendix Table 1). Though the 50% threshold is often still not met, the test regime under examination changes how far away the assessment is from the threshold. The adoption of more cognitively demanding assessments may make it easier to study HOTS going forward in states that are administering such exams today.

A second challenge is that higher-order thinking assessments are not typically administered longitudinally, limiting the ability to examine changes in higher-order skills or policy impacts over time. Assessments that do include a majority of items capturing HOTS such as portfolio- and performance-assessments (e.g., Koretz et al., 1996) or the Partnership for Assessment of Readiness for College and Careers (PARCC, Doorey and Polikoff, 2016) have been only briefly implemented. Researchers further document a range of measurement (Koretz et al., 1994), cost (Chingos, 2013), and political challenges (Jochim and McGuinn, 2016) that have since led to the discontinuation of dedicated K-12 HOTS assessments administered at scale. As a consequence, the ability for researchers to understand higher-order thinking outside of brief administration periods or the impacts of educational interventions on HOTS over time is limited. These collective challenges of largely no systematic, longitudinal administration of HOTS

assessments have led researchers to seek alternatives to using standalone assessment results to study HOTS.

Some scholars have attempted to measure higher-order skills using subsets of information available from large-scale assessments to draw conclusions about our educational systems, but the evidence for the validity of this approach is currently limited. For example, researchers have made use of information found in large-scale assessments to proxy for HOTS. These proxies include leveraging scores from pre-existing assessment subscales (e.g., Problem Solving) or generating scores for items already classified under the same topic/standard (e.g., Data Analysis) or item format (e.g., open-ended questions). Another possibility could be leveraging item difficulty as an indicator of whether an item captures higher-order thinking (Yuan and Engelhard, 2023). The appeal of such approaches lies in the ability to retroactively make use of pre-existing information commonly available in administrative datasets that could theoretically be applied to any national or state context over long periods of time. Using information found in large-scale assessments, scholars have asserted conclusions on several salient educational topics such as accountability policies (Jacob, 2005), teacher value-added (Lockwood et al., 2007; Papay, 2011), charter schools (Cohodes, 2016), extracurricular activities (Schueler & Larned, 2023) and educational inequality (Mitani, 2021). However, these strategies have tended to be deployed without an accompanying thorough examination of the validity of the measurement approach.

There are good reasons to suspect that the validity of such attempts to use a particular subscale, standard, or item format as a proxy for capturing HOTS should be carefully examined rather than taken as a given. One issue is that these proxies may come from assessments designed to capture a unidimensional construct, such as the PISA mathematics exam designed to assess mathematics literacy (e.g., Schleicher, 2024). This introduces a challenge of attempting to extract dimensionality from something designed to capture a single construct (Liu et al., 2018). Another potential challenge is that these proxies may not always meet the guidelines to justifiably capture HOTS. For example, Mitani (2021) uses the "Reasoning" subscale of the TIMSS, which consists of 20-25% of the assessment items (Philpot et al., 2021), as a measure of HOTS. However, Yuan and Le (2014) document that only 2% of TIMSS items capture HOTS: it is therefore not possible mathematically for the "Reasoning" subscale to be made up of a majority of items that capture HOTS[1]. Various standards or topics on four statewide ELA assessments that have been studied range from 0% to 67% (Wixson et al., 2002), and 0-6% on the ACT and SAT (Christopherson and Webb, 2018). Item format, particularly open-ended response items, may range from 0% capturing HOTS all the way to 100% but is completely dependent on the state and grade contexts (Yuan and Le, 2012, 2014; Christopherson and Webb, 2018). Therefore, there remains a need to examine evidence of validity for these approaches to capturing HOTS.

Examining a Subscore Approach to Capturing Higher-Order Thinking Skills at Scale

One related but underexamined approach to capturing higher-order skills at scale may resolve the previously outlined concerns: using assessments administered to broad, representative statewide populations, directly coding individual test items on whether they capture higher-order and lower-order thinking skills, and pairing the coding with student item-level responses to create subscores of both HOTS and LOTS. In essence, researchers could examine each item found on these large-scale assessments (e.g., statewide standardized exams

[1] If all 2% of TIMSS items were under the "Reasoning" subscale, this would leave the remaining "Reasoning" items as lower-order thinking skills. This would represent (2/25), or 8%, of the "Reasoning" subscale capturing HOTS.

used for accountability purposes in all states) and determine whether the items reflect lower- or higher-order thinking. Then, using data that contain students' item-level responses, researchers could compute how well students perform on higher-order items and lower-order items. This approach has the potential to address previously described challenges. The resulting HOTS and LOTS subscores that this approach would allow researchers to calculate would be available for statewide samples of students over long periods of time. Furthermore, if viable, such an approach could work in any context where both assessment item content and student performance on individual items are available—in theory, higher-order thinking could be examined in nearly any state or national context either in the present-day (Common Core-aligned assessments) or the past (No Child Left Behind-aligned assessments). Although the approach sounds appealing, scholars have not yet rigorously examined evidence for the validity of such an approach. For example, most achievement tests are assumed to be unidimensional and are specifically designed to assess a unidimensional construct (e.g., reading ability). Dividing such assessment items into HOTS and LOTS subscores could pose fundamental measurement issues, so a range of psychometric analyses are required to determine whether validity evidence supports their use for distinguishing HOTS from LOTS.

Therefore, in this paper, we investigate whether coding statewide assessment items on whether they capture higher-order skills to create a HOTS subscore can be supported by the psychometric evidence. We code Massachusetts's ELA assessment items from 2001 to 2023, including multiple testing regimes, and pair that coding with item-level student performance data to create a higher-order (and lower-order) thinking skills subscores. We examine whether supporting measurement validity evidence exists for the intended use of distinguishing HOTS from LOTS and generating knowledge about K-12 students' HOTS by: 1) establishing initial benchmarks of the strength of validity evidence for unidimensionality, 2) comparing measurement models designed to test different assumptions about items coded as capturing HOTS, and 3) examining characteristics of the best-fitting multidimensional measurement models. These efforts collectively allow us to address the following research questions:

1. What supporting validity evidence, if any, exists for the coding of state assessment items on whether they measure higher-order thinking skills and pairing that coding with performance data to capture higher-order thinking skills at scale? Does the evidence differ by test regime (i.e., pre- versus post-Common Core alignment)?
2. Does a higher-order thinking skills subscore generated through this process show additional desirable model characteristics, such as high factor loadings or the ability to predict later outcomes like educational attainment, that support its use for capturing higher-order skills?

Materials and Methods

Massachusetts Statewide Achievement, Attainment, And Demographic Data

We use publicly-available statewide data from Massachusetts on 3rd-8th & 10th grade student test item-level responses covering the 2000-01 to 2013-14 school years as well as the 2016-17 to 2022-23 school years.[2] We shorten school years to the spring year of the academic year for the remainder of the paper (e.g., we refer to the 2022-23 school year as 2023). The results are for the statewide exams used for school accountability purposes—the Massachusetts Comprehensive Assessment System (MCAS)—and therefore cover nearly all students in tested

[2] Massachusetts temporarily switched to the PARCC exam in the 2015 and 2016 school years, and the state does not publicly release those data. In addition, there is no testing data for the 2019-20 school year because of Covid-19.

grades in the state. We focus on results for the English-Language Arts (ELA) MCAS exams after finding that this subject had a greater concentration of items that capture higher-order thinking compared to other subjects (we describe this determination in more detail below). Data are at the student level but are anonymized and cannot be linked at the student level over time. These performance data are linked to information on the grade, year, modality (paper versus online), and session (students were randomly placed into different partially administered assessments groups in 2021 due to Covid-19). In total, there are 133 different assessment grade-year-modality/session combinations. The state provides the scoring for all of the possible responses to each item. Responses are ordinal categorical, featuring an "incorrect or correct" designation for the multiple choice items or a summarized judgment of a students' response (e.g., minimal understanding or exemplary understanding) for the open response items. The data contain over 200 million individual item responses and 9 million student-grade-year observations with approximately 70,000 students per grade-year combination (e.g., 3rd graders in 2001).

The data also includes information on school-level high school graduation and postsecondary enrollment rates for 9th grade cohorts from Spring 2003 through 2017[3]. While we are unable to link students to themselves across years, a public student-level dataset contains information on the school each student attends[4], which can be paired with public data on each school's annual graduation and postsecondary enrollment rates. The graduation data features information on the school's 9th grade cohort in a particular year (e.g., the 2003 cohort would start 9th grade in Spring 2003 and be expected to graduate high school in Spring 2006 if they progressed on time). The data includes 9th grade cohort student count and graduation rate within 4-years[5] after accounting for transfers in and out of the school[6]. The postsecondary enrollment data include the number and percentage of graduates that enrolled in higher education within 16 months (and whether it was in a 2- or 4-year and public or private institution). To create a cohort-based rate of postsecondary enrollment within 16 months (instead of just high school graduates), we divide the school's number of graduates that enroll in postsecondary institutions within 16 months by the school's number of students in the corresponding 9th grade cohort. Since 9th grade students do not typically take statewide ELA assessments, we link the graduation and postsecondary data to the school's equivalent 10th grade cohort (Spring 2004 10th grade cohort would use their school's Spring 2003 9th grade graduation and postsecondary enrollment rates).

A separate dataset also features demographic information available at the student level that contains the student-level performance data (but not the school a student attends). Table 1 describes the demographic breakdown across all years and over time. The racial composition of the sample is 6% Asian, 9% Black, 16% Hispanic, 66% White, and 3% as Native, Pacific Islander, or Mixed. The socioeconomic breakdown is 35% economically disadvantaged

[3] After this year, the public data changes to a different measure of high school graduates by March and is no longer comparable to previous years.

[4] This dataset is essentially identical to the previously described dataset except no demographic information is available. To preserve anonymity, Massachusetts offers two publicly available datasets: one with students' demographic information but no school- or district-level information, or one with students' school- and district-level information but no demographic information.

[5] Massachusetts also reports out a 5-year graduation rate.

[6] Massachusetts features a version of graduation rate that does not account for any transfers in and out of the school. Our preferred version is the one described in the text—4-year graduation rates after accounting for transfers in and out of the school—but our later analyses do robustness checks on all versions.

students[7]. The student gender breakdown is 51% Male and 49% Female, the special education population is 18%, and the English Language Learner composition is 7%. The percentage of non-White, economically disadvantaged, special education status, and English Language Learner students gradually increased over the two decades in Massachusetts.

Higher-Order (and Lower-Order) Thinking Skills Data

Coding Process

We merge the statewide data on test item performance with an original dataset for which we manually coded whether or not each ELA test item captures higher-order thinking skills. Webb's (2002) Depth of Knowledge (DOK) framework forms the basis for our coding. While alternative frameworks (e.g., Bloom's Taxonomy) exist for conceptualizing and operationalizing higher-order thinking, Webb's framework is appealing from a policy researcher perspective. DOK is a common method of assessing cognitive demand (Martone and Sireci, 2009) that is still used in designing assessments across multiple large-scale contexts, ranging from 26 states' accountability exams[8] (Appendix Table 1) to NWEA's Measures of Academic Progress (MAP) Growth. Furthermore, scholars applied DOK as the key criterion—50% or more of an assessment should consist of items that measure HOTS—for determining whether assessments adequately captured higher-order thinking skills (Darling-Hammond et al., 2013). Beyond the broad implications outside of Massachusetts, the framework closely maps onto our conception of higher-order thinking skills centered on analytical, reasoning, and problem-solving skills. DOK includes the following four levels:

- DOK 1 - Recall/Reproduction: "Recall of a fact, term, principle, concept, or perform a routine procedure."
- DOK 2 - Skills/Concepts: "Use of information, conceptual knowledge, select appropriate procedures for a task, two or more steps with decision points along the way, routine problems, organize/display data, interpret/use simple graphs."
- DOK 3 - Strategic Thinking: "Requires reasoning, developing a plan or sequence of steps to approach problem; requires some decision making and justification; abstract, complex, or non-routine; often more than one possible answer"

[7] The definition of economically disadvantaged has changed over time in Massachusetts. The original definition from 2001 to 2014 consisted of whether a student was eligible for free-reduced price lunch. In the 2011-2012 school year, the U.S. Department of Agriculture introduced the Community Eligibility Provision, which allowed all students in schools/districts with high concentrations of low income students to receive free meals—as a consequence, data on free-reduced lunch eligibility became less available as more schools and districts adopted the Community Eligibility Provision. In the 2014-2015 school year, the state shifted to "economically disadvantaged," which flags a student as economically disadvantaged if the student participates in the Supplemental Nutrition Program, the Transitional Assistance for Families with Dependent Children, the Department of Children and Families' foster care program, or Medicaid. Finally, on top of the previously mentioned components of economically disadvantaged, starting in the 2021-2022 school year, Massachusetts further added a supplemental data collection form to identify students that meet the 185% of federal poverty level threshold but were not identified under the previous criteria. Massachusetts also included students that were reported by the district as "homeless" as part of their economically disadvantaged definition.
https://www.doe.mass.edu/infoservices/data/ed.html (2001-2021 Information).
https://www.doe.mass.edu/infoservices/data/sims/redefining-lowincome.html (2022-2023 Information).

[8] We examined the test blueprints for each state context (or assessment consortium), documenting whether DOK was embedded in the design of the assessment.

- DOK 4 - Extended Thinking: "An investigation or application to real world; requires time to research, problem solve, and process multiple conditions of the problem or task; non-routine manipulations, across disciplines/content areas/multiple sources"

DOK serves as a useful approach because the framework describes the type of thinking required of students to interact with a particular test item. DOK levels 3 and 4 both encompass key features of higher-order thinking (e.g., "reasoning," "requires time to research, problem solve"). Therefore, we coded each individual ELA test item on a 1-4 DOK scale, flagging items with a rating of 3 or 4 as likely to capture higher-order thinking. We coded a total of 3,570 ELA assessment items. This represented the full universe of ELA items administered from 2001 to 2023 for which Massachusetts made the text of the items themselves publicly available. The item text was always available from 2001 to 2008 and for 10th grade assessments but not always for other grade-years. Overall, 46% (1,835/3,977) of 3rd-8th grade assessment items between 2009 to 2023 were not released publicly, so they were not coded here—we tackle this issue more in-depth during our validity checks but conclude that released items are representative of the full universe of ELA items on numerous observable dimensions.

To determine item DOK levels, the coders read through the text of each item carefully and leveraged Hess's cognitive rigor matrix (2009), which provides a detailed set of descriptions of items at each DOK level, to locate an item description that matched the item and then record the corresponding DOK level. We provide this matrix in Appendix Figure 1. We also provide illustrative contrasting example items. Appendix Figure 2 contains a reference to a passage as well as one item we coded as DOK level 1 (top) because it asks students to apply standardized rules to identify the part of speech, and another item we coded as DOK level 3 (bottom) due to the requirement students provide a characterization with supporting textual evidence. Appendix Figure 3 shows a completely different passage reference with one item we coded as DOK level 2 (top) because it asks students to summarize the information, and another item we coded as DOK 3 (bottom) because the item asks to analyze the author's craft, specifically the rationale for the use of a literary device.

A trained research assistant (RA) double coded items with a lead coder—the first author of this paper, who brought previous training in DOK coding as a K-12 teacher along with topical expertise—until the RA and lead coder achieved a 90% agreement rate across 267 items from 10 assessment combinations in different grades and years. The RA then coded the rest of all the items, and the lead author randomly double-coded 33% of those remaining items. The team met weekly to resolve questions and discrepancies in coding, formalize decisions into a coding guide document, and if needed, escalate questions or discrepancies to the second author for resolution. The overall agreement rate was 90.6% and even higher for whether items were classified as DOK 3 or 4 (HOTS) at 95.5%. The final coding rate of items capturing HOTS aligns with findings from other scholars that coded a subset of Massachusetts's ELA assessments for DOK (Yuan and Le, 2012; Doorey and Polikoff, 2016).

Coding Results

Our item coding revealed that these pre-existing assessments across both grades and years possessed sufficient variation in the extent to which they captured higher-order (versus lower-order) thinking for us to further examine validity evidence for using our approach to measure HOTS and compare them with LOTS. Table 2 documents the results of the coding for each grade over time. Every grade-year combination included at least some items that captured HOTS, but no assessment combination had at least half of the items classified as HOTS. We

show that across all grades and years, an average of 19% of items were coded as HOTS (sum of items coded DOK 3 or DOK 4). The grade-by-grade average ranged from 9% to 24% of items coded as HOTS, and the year-to-year average ranged from 11% to 35% of items coded as HOTS. As a reminder, we do not observe data from the Common Core-aligned PARCC assessments that were administered in two years between the MCAS 1.0 era and the transition to MCAS 2.0. However, we do observe multiple years in both the MCAS 1.0 and MCAS 2.0 periods. We document a substantial increase in the share of items assessing HOTS that emerged with the introduction of the MCAS 2.0 (Appendix Figure 4). Table 3 shows that the percentage of HOTS items in MCAS 1.0 is 17% but jumps up to 28% in MCAS 2.0; the percentage of HOTS points on the assessment similarly increase from 35% to 53% after the shift. The change happened across all grades. The shift came from a reduction of DOK 1 items: an example is shown in Appendix Figure 5 with DOK 1 making up 21% of assessment items in 2001 (MCAS 1.0) but only 5% in 2023 (MCAS 2.0).

Testing Various Measurement Model Assumptions

To explore whether there is evidence for the validity of this subscale approach to capturing HOTS and distinguishing them from LOTS, we begin by fitting measurement models using confirmatory factor analysis and assessing model fit indicators. This will help us understand whether models that align with the assumption that the items we have flagged as capturing HOTS indeed seem to measure a distinct construct or sub-construct from the items flagged as capturing LOTS (i.e., that the data are not purely unidimensional). The first measurement model is a baseline unidimensional model to which we will compare our multidimensional models. This first model assumes that the data are unidimensional, in other words, that a single ELA ability factor adequately fits the item response data. In Figure 1, we provide a graphical representation of (a) the unidimensional model in which all items, regardless of whether our team coded them as capturing HOTS or LOTS, serve as indicators of ELA ability. We anticipate this model will fit the data well given the state built the MCAS with unidimensionality in mind and the goal of capturing a single latent ELA construct.

In Figure 1, we also provide a graphical representation of (b) the bidimensional model. This model assumes that our item coding allows for differentiation between HOTS and LOTS. The bidimensional model, sometimes called a correlated factors model, no longer includes a single ELA factor but instead features two distinct but related latent factors—one for higher-order thinking skills and another for lower-order skills. Items coded as capturing HOTS are allowed to load onto the HOTS latent construct, while the LOTS items load onto the LOTS construct. The two latent factors (HOTS and LOTS) are allowed to covary, aligning with the theoretical conceptualizations of higher-order and lower-order thinking as skills within a broader domain (ELA in this case) that may depend on each other (e.g., LOTS development may be foundational for HOTS development). If the bidimensional model fits the data well, this would provide evidence that the HOTS and LOTS sets of items represent related but separate subscales that capture additional signal beyond a single overall factor representing ELA skills.

The final model from Figure 1 we investigate is (c) the testlet model, a special case of the bifactor model (Rijmen, 2010). A bifactor model assumes that all items do indeed measure general ELA ability, but that there could also be lingering variation in responses for items that capture higher-order and lower-order thinking skills beyond the overall ELA factor. For example, achievement test developers often assume that math is unidimensional, but that there is nonetheless meaningful variation at the subtest level shared by items related to specific content

like fractions, decimals, negative numbers, etc. The bifactor model mirrors this assumption, allowing for a general factor that cuts across all items, as well as specific factors corresponding to subtests/skillsets like fractions. If the bifactor model fit the data well, this would be consistent with a conceptualization of HOTS as domain neutral. In other words, that HOTS is a generalizable skill not specific to academic subject (e.g., ELA, science, history) and therefore applicable to any domain.

Here, we mainly focus on a special case of the bifactor model that may better align with our conceptualization of HOTS—the testlet model, which constrains item parameters of each specific factor (HOTS and LOTS) to be proportional if not equal to those from the general ELA factor for a given item (Li et al., 2006). That is, item parameters are invariant across dimensions. For example, the loading for the first item would be the same for the general (ELA) and specific (HOTS and LOTS) factors. In essence, the testlet model presumes each HOTS (or LOTS) item equally reflects HOTS (or LOTS) and a general ELA factor.

We employ confirmatory factor analysis (CFA) to fit each of the three measurement models separately for each of the 133 assessment combinations (by grade-year-modality/session). Since we have categorical data, we employ diagonally weighted least squares as our estimator and obtain robust standard errors along with mean- and variance-adjusted test statistics through weighted least square mean and variance adjustments. We first examine the strength of evidence supporting a unidimensional measurement model for Massachusetts's ELA assessment. To start, we establish whether the unidimensional measurement model is considered a "good fit" through the use of common fit statistics. Using the estimated measurement models, we extract the Root Mean Squared Error of Approximation (RMSEA), Standardized Root Mean Squared Residuals (SRMR), Comparative Fit Index (CFI), and the Tucker-Lewis Index (TLI). We then employ traditional cut-offs for each RMSEA, SRMR, CFI, and TLI (e.g., RMSEA $<$ 0.05 and CFI $>$ 0.95) following Hu and Bentler (1999) to make that determination. Next, we examine whether either of our multidimensional measurement models provide evidence either in favor or against unidimensionality. If the correlation between HOTS and LOTS factors in the bidimensional model is essentially one, then the unidimensional and bidimensional model are essentially identical, suggesting unidimensionality. Similarly, if little to no meaningful variation for HOTS or LOTS remains after accounting for the general ELA factor in the testlet model, this would also suggest that the assessment captures a unidimensional construct. However, if correlations between HOTS and LOTS are not high with the bidimensional model or there is non-trivial lingering HOTS and LOTS variation with the testlet model, then multidimensional measurement models could potentially be supported.

Additional Psychometric Analyses for Assessing the Validity Evidence

We also identify additional sources of model validity evidence for the intended use of generating knowledge about HOTS for research, policy, and practice that could further substantiate the use of multidimensional models. These consist of model comparisons, factor loadings, and whether subscores predict future outcomes that we would expect to be correlated with HOTS. Comparing fit statistics across models could help provide evidence of the relative fit and therefore appropriateness of the unidimensional versus multidimensional models. While we might anticipate the model fit to improve with additional complexity, superior fit statistics across all RMSEA, SRMR, CFI, and TLI (i.e., lower RMSEA and SRMR as well as higher CFI and TLI), which account for model complexity via RMSEA and TLI (e.g., Hu and Bentler, 1999), would show supplementary evidence that one measurement model could better encapsulate the

underlying data structure. Similarly, reasonably high factor loadings would provide additional evidence in support of a given measurement model and the theoretical conception is represents. If the factor loadings of the model are relatively high, this would suggest that a strong relationship exists between the latent factor and the items. Conversely, if the factor loadings are extremely low, then perhaps the relationship may not be fully supported.

Finally, the extent that students' subscores on higher-order thinking skills predict later life outcomes could be a meaningful piece of evidence. If the subscore positively predicts later life outcomes, this provides further evidence consistent with the idea that our subscore is functioning as intended given the prior research connecting HOTS skills and postsecondary educational attainment (Facione, 1990; Zahner & James, 2015). To examine whether students' subscores on our HOTS measure predict later life outcomes, we generate HOTS and LOTS factor subscores for our multidimensional measurement models. Since we cannot link students and outcomes, we must use 10th grade cohorts. As described in the data section, students in 10th grade can be linked to their school-level graduation and postsecondary outcomes. We generate factor scores estimated with the default approach for categorical data—empirical Bayesian modal—for 10th graders. These factor subscores are standardized within cohort[9]. To study the association between standardized scores for individual students and school-level postsecondary outcomes, we estimate the following linear regression:

$$Y_{ts} = \beta_o + \beta_1(HOTS\ Subscores_{it}) + \beta_2(LOTS\ Subscores_{it}) + \delta_t + \varepsilon_{its} \quad (1)$$

Y_{st} represents the outcome of interest (e.g., 4-year graduation rate) for cohort *t* in school s. β_o is the intercept, δ_t consists of cohort fixed effects to ensure outcome comparisons are made only within groups of students that start 10th grade at the same time[10], and the ε_{its} represents the error term clustered at the school-level. We also incorporate student's lower-order thinking, β_2, in the model to control for those skills' association with educational attainment and isolate the unique relationship between HOTS and the outcomes. β_1 represents the coefficient of interest using student *i*'s standardized HOTS subscore in cohort *t*. If the coefficient is positive and statistically significant, this aligns with the hypothesis that HOTS positively predict educational attainment outcomes, reflecting that the subscore appears to capture meaningful HOTS skills beyond the aspects connected to LOTS.

Results

Leveraging Extant Statewide Data to Distinguish Higher- from Lower-Order Skills Is Challenging, Though Not Impossible Depending On Test Regime

Massachusetts's unidimensional measurement model shows supporting validity across test regimes (pre- and post-Common Core alignment), which is not surprising given these tests were designed to capture a unidimensional construct of ELA ability. Table 4 describes the fit of the unidimensional measurement model to the data across all assessment combinations (grade-year-modality/session), test regime (MCAS 1.0 versus MCAS 2.0), and grade. Fit statistics for each assessment combination across measurement models can be found in Appendix Table 2. The unidimensional measurement model specification, regardless of whether we use all

[9] Technically since our year range is from 2004-2018 for 10th graders, there is no need to standardize within modality or session.

[10] Since we separately estimate each assessment combination, the assessments (and characteristics) are already held constant. Aspects like the item difficulty are already accounted for because this analysis exclusively leverages student-level HOTS and LOTS variation to predict school-level educational attainment outcomes, minimizing those validity threats.

assessment combinations or only released ones, meets Hu's and Bentler's (1999) good fit criteria in 100% of the 133 assessment combinations. When looking across test regimes, the unidimensional model still fits the data well—MCAS 1.0 shows only slightly worse but still acceptable fit statistics (e.g., average SRMR of 0.032) compared to MCAS 2.0 (e.g., average SRMR of 0.029). Since the evidence largely suggests that unidimensional assumptions are supported, attempting to capture HOTS with such extant data appears challenging.

In particular, the evidence does not support using the Common Core-aligned MCAS 2.0 assessments to distinguish HOTS from LOTS. Multidimensional measurement models incorporating information on HOTS nearly always show evidence that the assessment is unidimensional under MCAS 2.0. To illustrate this more concretely, Table 5 presents the full parameterization of the bidimensional and testlet models for two example assessment combinations, one under MCAS 1.0 (10^{th} graders in 2014) and another under MCAS 2.0 (10^{th} graders in 2023). The MCAS 2.0 bidimensional model shows an extremely high correlation of 0.96 between the HOTS and LOTS factors, signaling that the model largely reflects a unidimensional construct. This is consistent across assessment combinations. In Figure 2, we show the correlation between HOTS and LOTS factors for the bidimensional models that converged (100 of 133 grade-year combinations) by testing regime. Correlations between HOTS and LOTS factors with MCAS 2.0 data feature extremely high values ranging from 0.925 to 1.000, suggesting that the assessment largely represents a unidimensional construct.

The testlet model results also largely point to a unidimensional construct. After accounting for the general ELA factor, the testlet model with MCAS 2.0 data simply does not converge (estimating impossible negative HOTS variances). This is again consistent with the idea of unidimensional MCAS 2.0 assessments. Figure 3 plots the magnitudes of the general ELA factor and either the HOTS or LOTS variance (whichever is smallest) for the converging (31/133) and non-converging (100/133) testlet models. After accounting for the general ELA factor, the only two converging testlet models under MCAS 2.0 leave almost no lingering HOTS/LOTS variance (nearly zero), suggesting that the assessment measures a unidimensional construct instead of capturing distinct HOTS/LOTS dimensions. Indeed, we show in Appendix Figures 6, 7, and 8 that the little lingering HOTS or LOTS variance is associated with stronger evidence of unidimensionality. For example, the assessments for which we observe less lingering HOTS and LOTS variance using the testlet model are the assessments with higher correlations between the HOTS and LOTS factors when using the bidimensional model[11]. Altogether, the

[11] We explore potential explanations for the little variation by associating the smallest magnitude between either the HOTS or LOTS variance and characteristics of alternative measurement models (i.e., unidimensional model's fit statistics and bidimensional model's HOTS and LOTS factor correlations). Appendix Figures 6, 7, and 8 show scatterplots from that exploration. We find that the stronger the case for unidimensionality—lower unidimensional RMSEA and higher unidimensional TLI plus higher bidimensional HOTS and LOTS factor correlations—the magnitude of the converging testlet model's HOTS or LOTS variance tends to be smaller. For example, the correlation between the bidimensional model's HOTS and LOTS factor correlation and smallest of testlet model's HOTS or LOTS variance is -0.70, suggesting a strong, negative association between the two.

We also investigate non-converging models, finding some additional, albeit imperfect, evidence that especially strong evidence in favor of unidimensionality limits the lingering HOTS or LOTS variation in testlet models. Although Appendix Figures 6, 7, and 8 show negligible correlations of -0.22 to 0.18 between non-converging testlet model's smallest HOTS or LOTS variance and unidimensional and bidimensional model characteristics, differences emerge when separating out HOTS and LOTS variance. Appendix Figures 9 to 14 reproduce Appendix Figures 6, 7, and 8 for the HOTS and LOTS variance. We find that the relationship between stronger unidimensional models and smaller variance better holds for HOTS variance (e.g., -0.58 correlation between HOTS variance and bidimensional

findings for these multidimensional measurement models suggest that the MCAS 2.0 assessments are not a good source of information for distinguishing HOTS from LOTS.

Pre-Common Core MCAS 1.0 assessments are also not often a good source of information for capturing HOTS, however, some grade-year results do appear to allow for differentiation between HOTS and LOTS. In many cases, we observe non-trivial magnitudes of HOTS or LOTS variance left after accounting for the general ELA factor in the testlet model with MCAS 1.0 data. Figure 3 shows that the lingering HOTS or LOTS variance magnitudes typically remain below 0.10, suggesting that, with the testlet approach, we are often trying to treat a unidimensional construct as something it's not. That said, the testlet model does converge in a nontrivial number of cases (29), and the bidimensional models in MCAS 1.0 sometimes show correlations low enough (average of 0.86) to warrant exploring their use in differentiating between HOTS and LOTS, especially compared to MCAS 2.0, which had an average correlation of 0.97. While the correlations for MCAS 1.0 range from as low as 0.66 to as high as 0.99, as seen in Figure 2, across almost all grades, Table 6 documents that 75-100% of converging bidimensional model's HOTS and LOTS correlations in MCAS 1.0 fall below 0.90 with 17% below 0.80. The only exception is 3rd grade, which has a higher average HOTS and LOTS correlation of 0.91. For some of the assessment combinations, like the MCAS 1.0 example in Table 5 (visualized in Figure 4), differentiation between HOTS and LOTS appears to be possible. This evidence, in conjunction with additional psychometric analyses, could support the use of extant item-level information for distinguishing HOTS from LOTS when using certain grade-years from the MCAS 1.0 assessments.

Additional Analyses Support the Use of Certain Grade-Years of Pre-Common Core-Aligned Assessments For Distinguishing Higher- from Lower-Order Skills

The bidimensional modelfits the MCAS 1.0 data well in the majority of cases, with fit statistics that are virtually indistinguishable from an already well-fitting unidimensional measurement model. Table 7 summarizes the total number of MCAS 1.0 assessment combinations, that converged with bidimensional models, and whether the models meet Hu and Bentler (1999) criteria. The bidimensional model, like its unidimensional counterpart, meets the Hu and Bentler (1999) good fit criteria for all 100 models that converged. Furthermore, Table 7 reports out the percentage of grade-year-modality-session combinations that show a superior fit (lower RMSEA/SRMR and higher CFI/TLI) over the unidimensional models among converged bidimensional models. Across unidimensional model comparisons used, the fit statistics for the bidimensional model are better than the fit statistics of the unidimensional model in the majority of cases (78%) for MCAS 1.0. That said, the unidimensional models already achieve essentially perfect fit, so we basically interpret the unidimensional and bidimensional models as having indistinguishable fit. One potential concern is that the slightly better, smaller RMSEA estimates of the bidimensional model are not dramatically different from the unidimensional model's RMSEA. However, a more conservative comparison using RMSEA confidence intervals, specifically a smaller RMSEA for the unidimensional model and a larger RMSEA for the

---

model's HOTs and LOTS factor correlations) but not LOTS (i.e., correlations of nearly zero). Other factors—perhaps too many HOTS items leave too little LOTS variation or maybe the larger role of the General ELA factor minimizes the lingering role of LOTS—also do not explain the minimal LOTS variance, as seen in Appendix Figures 15 to 18. Though we are unable to draw fully concrete conclusions, especially with respect to LOTS variance, we see some signs that point us toward the strength of evidence for unidimensionality associated with testlet models leaving little lingering HOTS or LOTS variance.

bidimensional model, mostly affirm the results. In Figure 5, we plot the lower 90% confidence interval of RMSEA for the unidimensional models compared to the upper 90% confidence interval of RMSEA for the converged bidimensional models[12]. All grades except 3rd still show superior fit statistics over the unidimensional model in the more conservative comparison, presenting initial validity evidence that supports the ability to differentiate higher-order and lower-order thinking skills when using MCAS 1.0[13]. To be clear, although we observe slight differences in fit statistics between the bidimensional and unidimensional model, the two both fit the data well, limiting the degree we can make distinctions between the two models.

Further examination of bidimensional model characteristics highlights additional evidence potentially supporting their use with MCAS 1.0 data. For instance, the bidimensional model's factor loadings for both HOTS and LOTS appear fairly high. Table 8 provides the *average* minimum, mean, standard deviation, and maximum standardized factor loadings for HOTS and LOTS across MCAS 1.0 grades. Factor loadings for both HOTS and LOTS across MCAS 1.0 years and grades average at relatively high values of 0.66 and 0.58 respectively. Individual grades under MCAS 1.0 show similar average factor loadings of 0.62 to 0.69 for HOTS and 0.55 to 0.61 for LOTS. Altogether, these results suggest a moderate-to-strong relationship between the latent factors and their items that support the potential use of the bidimensional model to capture higher-order skill signal and distinguish it from LOTS.

Higher-Order Skills Subscores Predict Educational Attainment Outcomes

We also examine the HOTS subscale's ability to predict later educational attainment outcomes. Students' HOTS positively predict their school-level high school graduation and postsecondary enrollment rates. Table 9 reports out the linear regression estimates between students' K-12 skills and school-level educational attainment outcomes. Explicitly including students' lower-order thinking subscores as a predictor in the regression model addresses concerns that observed associations between HOTS and outcomes of interest may simply reflect LOTS. When we control for LOTS, the relationship between HOTS and high school graduation remains positive in direction but is no longer statistically significant. Therefore, K-12 LOTS skills appear more predictive of graduation than HOTS skills. A different pattern emerges when it comes to postsecondary attainment, where HOTS remains predictive of enrollment in postsecondary institutions within 16 months of graduation even after accounting for their LOTS. Specifically, we find that a one standard deviation (SD) increase in our K-12 HOTS subscore is associated, on average, with a statistically significant 1.26 percentage point increase in the school's postsecondary enrollment rate. Interestingly, the magnitude of the relationship between

[12] We also show the fit statistic comparisons in Appendix Table 3.

[13] Comparing measurement models with approaches beyond changes in fit statistics affirms that the bidimensional model fits the data similarly to the unidimensional model in MCAS 1.0. Reaching similar conclusions about bidimensional fit over unidimensional models with different measurement comparison approaches would reassure us that the results are robust. We compare bidimensional and unidimensional model fits using chi-squared difference tests as well as approaches that treat the binary and polytomous items as continuous instead of categorical. In Appendix Table 4, we show that every converging assessment combination of the bidimensional model is significantly better than the unidimensional model on the chi-square statistic. Similarly, in Appendix Tables 5 and 6, we find that across both Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Adjusted BIC as well as Likelihood Ratio Tests, the bidimensional model continues to provide a superior fit over the unidimensional measurement model (over 89% of assessment combinations). These results suggest that differentiating higher-order and lower-order thinking skills can be supported in MCAS 1.0.

LOTS and postsecondary enrollment rates is larger (5.36 percentage points) than the magnitude of the relationship for HOTS.

The bottom panel of Table 9 further shows that the positive association between HOTS and postsecondary enrollment is driven by 4-year postsecondary enrollment rather than 2-year postsecondary enrollment. On average, a 1 SD increase in HOTS is associated with a 1.46 percentage point increase in the school's percentage of students enrolling in 4-year postsecondary institutions within 16 months of graduation after conditioning for a student's LOTS[14]. These postsecondary enrollment magnitudes are non-trivial—with the average postsecondary enrollment at 67% and 4-year postsecondary enrollment at ~49% across all years, the coefficients reflect increases of 1.9% and 3.0% respectively compared to the overall sample average. Appendix Table 11 provides a breakdown for each year, documenting that these results are largely positive across individual cohorts. Altogether, these results are consistent with the idea that the subscale captures some additional signal beyond general ELA ability.

Validity Checks

To review, we find mixed evidence on the viability of the subscale approach to capturing higher-order skills and distinguishing them from lower-order skills—that is coding extant statewide test items based on DOK and linking the coding to performance data. More specifically, we find evidence suggesting that this is not an appropriate approach with data from the MCAS 2.0 era in Massachusetts and for many of the grade-year cases in the MCAS 1.0 era. Therefore, we caution against applying this approach in most contexts. That said, we do find evidence that it is possible to use this approach in certain grade-year cases in the MCAS 1.0 era and therefore some subset of grade-year exam results could be used to learn more about higher-order skills if researchers carefully examine the measurement properties of the data prior to applying it to this end. For researchers who may be interested in using the subscale approach where it is possible—for example, in certain grade-year cases in the MCAS 1.0 era—we provide further evidence of validity for this approach below.

Released Assessment Items Are Representative of The Full Universe of Items

[14] Higher-order skills remain predictive of postsecondary enrollment across numerous different outcome definitions. Another threat to the predictive validity evidence would exist if the results changed when different outcome definitions are used. Appendix Table 7 replicates Table 9 using three alternative outcome definitions: 1) 4-year graduation and postsecondary enrollment rates that accounts for students transferring out but not students transferring in, 2) a 5-year graduation and postsecondary enrollment rate that accounts for students transferring out or in, and 3) a 5-year graduation and postsecondary enrollment rate that accounts for students transferring out but not students transferring in. Across these definitions, HOTS positively predict these postsecondary outcomes, including postsecondary enrollment after accounting for other skills for the bidimensional model. Like before, this result also holds for 4-year postsecondary institutions as seen in Appendix Table 8. Finally, we also use the original school's percentage of high school graduates enrolling in postsecondary institutions within 16 months provided by Massachusetts in Appendix Table 9 and find a significant positive association for HOTS.

The final predictive validity check aggregates students' higher-order thinking (and lower-order thinking) subscores to the school-level: higher-order scores are still associated with positive educational attainment. Since the educational attainment outcomes are at the school-level, one threat could be the use of individual HOTS subscores to make predictions on school-level outcomes. Appendix Table 10 shows the results of Equation 1 after obtaining the average HOTS and LOTS subscores from the bidimensional model for the school. Again, HOTS remain predictive of postsecondary enrollment, particularly 4-year postsecondary enrollment, even after accounting for students' LOTS, suggesting that the HOTS subscale is meaningful.

One potential threat to the validity of the HOTS measure relates to the partial public release of assessment item content. Although student item-level performance data is available for the entire period of study, the assessment item content (i.e., the text of the test questions themselves) is only fully released from the 2001 to 2008 school years and for 10th graders. For other grade-year combinations, we have a partial set of the items. For 3rd-8th graders in 2009 and beyond, the content of assessment items is released for 2,142 of 3,977 items (or 54%). Since the content of the assessment item determines its DOK coding, one concern could be that the selective release of assessment items biases results. Perhaps in the 2009-2023 period Massachusetts released a disproportionate share of HOTS vs. LOTS items. This could happen, for example, if they released only open-ended response questions to allow teachers to prepare students for those types of questions or if the test designers prioritized releasing the easiest questions to alleviate test-taking anxiety. This could threaten the validity of our conclusion that this approach to capturing HOTS could be applied across contexts (with a full universe of assessment items).

Therefore, we examine the representativeness of released assessment items compared to the full universe of test questions based on the items' corresponding subscale, format, and difficulty. First, Massachusetts data provides a corresponding subscale for each item. The Language subscale encompasses language convention and writing, while the Reading subscale involves items that capture text comprehension. Each item is also categorized as either multiple choice or some form of open response (e.g., short response or essay). In addition, students' item-level responses allow us to generate item difficulties for each item by replicating Massachusetts's measurement model used for the assessments (Item Response Theory Three-Parameter Logistic and Graded Response Model). The mean difficulty is used for each polytomous item, which is allowable in cases that have no threshold reversals and monotonically increasing thresholds such as ours (Ali et al., 2015). Using these item characteristics, we document the average subscale, item format, and item difficulty for all items versus the subset of released items used for our HOTS measure. We also estimate a linear (and additional logistic for binary outcomes) regression to examine the relationship between the item characteristic and the likelihood of item release, while including grade-year-modality fixed effects to ensure assessments are only compared to themselves.

Analyses indicate that the released items are representative of the full universe of items on subscales, item formats, and item difficulties. Appendix Table 12 reports the average percentage under each subscale and item format along with the average difficulty for all items and released ones. The differences between all items and released ones are small: 3% for subscales, 1% for item formats, and 0.15 for item difficulty. The subsequent regression models in Appendix Tables 13, 14, and 15 show a consistent null association. No statistically significant difference exists between released items and unreleased items on subscales and item formats across every single year and grade. The same largely holds for item difficulty with no statistically significant difference between released and unreleased items except for in a few years (2014, 2017, and 2023). However, since the statistically significant coefficients are both positive and negative in those years, the possibility of systematic selection bias seems less likely. These analyses altogether reassure us that the representativeness of released assessment items likely does not pose a validity threat.

DOK Coding Captures HOTS Variation Beyond Other Item Characteristics

Another potential concern is that our depth of knowledge coding conflates the higher-order and lower-order thinking constructs with other item characteristics. For example, if all the items coded as higher-order thinking also were simply the most difficult items on the test, this would limit the ability to claim we are capturing students' higher-order thinking. Instead, we may be detecting students' (in)ability to answer the most difficult questions. Another implication would be that it may not be worth the time and effort to code individual items based on the DOK framework. A mixture of both lower and higher difficulty HOTS items would lessen concerns that the construct could be conflated with other item characteristics. To examine the extent to which our depth of knowledge coding reflects other item characteristics, we document the share of HOTS and LOTS items that fall under each subscale and item format, and plot the distribution of item difficulty, by test regime and grade.

The evidence suggests that HOTS items tend to exhibit certain item characteristics but cannot be reduced to any particular item characteristic. In Appendix Table 16, we identify the percentage of HOTS and LOTS items for the Language vs. Reading subscale as well as the multiple choice vs. open-response item formats. HOTS items are slightly more likely to be part of the Reading subscale (89%) and far more likely to be open-response (47%) compared to LOTS items at 81% and 0%, respectively. However, a non-trivial share of HOTS items are multiple choice (53%), suggesting that our HOTS measure is not simply capturing whether an item is open-response. We also display a boxplot of the difficulty distribution of HOTS and LOTS items in Appendix Figure 19. HOTS items tend to be slightly more difficult compared to LOTS items, but significant heterogeneity exists among HOTS items. As we show in Appendix Figure 20, some grades, like 3rd, nearly always have HOTS items far more difficult than LOTS items, while other grades like 10th sometimes suggest HOTS and LOTS items have similar difficulties. In sum, although HOTS coding is correlated with item characteristics, no single item characteristic fully captures higher-order thinking, signifying that the DOK coding does capture some degree of HOTS variation beyond the other characteristics of exam questions. Researchers interested in leveraging this depth of knowledge coding scheme could re-estimate results separately by item characteristics (e.g., using only HOTS items that are multiple choice or similar difficulty to LOTS) as a robustness check to lessen concerns of conflating HOTS with other item characteristics.

HOTS Measures Can Be Used to Make Subgroup Comparisons

One final desirable property that is especially useful for the examination of educational inequality is the demonstration of scalar measurement invariance. Scalar measurement invariance allows for between-group average comparisons (Putnick and Bornstein, 2016). Using multiple group confirmatory factor analysis, for the bidimensional models, we show in Appendix Table 21[15] that group average comparisons on the basis of race and economic disadvantage are possible in over 95% of grade-year combinations using standard cut-offs (Chen, 2007)[16]. The exact grade-years slightly differ depending on race or socioeconomic status.

---

[15] The results for any individual grade-year on measurement invariance (factor, metric, and scalar) can be found in Appendix Table 1 and Appendix Table 2.

[16] We prioritize the measurement invariance results using alternative fit indices. Another approach for measurement invariance involves the use of absolute fit indices (i.e., whether there is a significant change in the chi-squared). Though Putnick and Bornstein (2016) recommend that both approaches be documented, it is also important to note that absolute fit tests of measurement invariance are sensitive to small, unimportant deviations from the perfect model under cases with larger sample sizes (Cheung and Rensvold, 2002). The extremely large sample sizes in this dataset mean that measurement invariance will nearly always be rejected for potentially unimportant deviations

LOTS and HOTS Largely Appear to Develop Sequentially

If indeed the bidimensional model was capturing distinct HOTS and LOTS skills, one expectation we would have is that we would observe evidence that LOTS and HOTS develop sequentially, i.e., that mastery of higher-order thinking skills first requires mastery of foundational domain knowledge (LOTS). We might anticipate that students must first master certain lower-order skills before being able to engage in higher-order thinking. For example, a student may need to know the definition of a theme (lower-order skill) before being able to provide textual evidence to support the identification of a theme (higher-order skill). It would be surprising if students were frequently able to do the reverse, providing textual evidence for a theme without even knowing what a theme is. Understanding whether the sequential hypothesis is supported empirically would therefore provide further reassurance that the HOTS coding, at least for the MCAS 1.0 data, allows us to capture HOTS and distinguish it from LOTS.

To test this sequential hypothesis, we begin by identifying the higher-order and lower-order skill items that assess the same ELA standard. We would not expect HOTS to be dependent on LOTS across ELA standards that assess very different skills[17]. To build on our previous theme example, we might not expect the ability to provide textual evidence to support the identification of a theme to be dependent on knowing grammar conventions (another lower-order thinking skill). Instead, we might expect HOTS items under the same standard as LOTS items to be potentially dependent on each other. In Appendix Table 17, we document the number of standards by test regime and grade along with the percentage of those standards that have only LOTS items, HOTS items, or a combination of the two. Our table shows that nearly every standard is composed either exclusively of LOTS items or a combination of LOTS and HOTS items. Only a few standards feature only HOTS items, but after we restrict the sample to standards with five or more released items[18], we find that this appears to be largely an artifact of standards with few released items. There are ample ELA standards under which we find both LOTS and HOTS items, signaling potential dependencies between the two.

When we examine student performance on higher-order and lower-order items within the same ELA standard, we find that it is rare for students to answer HOTS items correctly without answering LOTS items correctly, supporting the sequential hypothesis. We show this after limiting the sample to 10th grade students (since all assessment items are released for this group). We subsequently compute 10th grade students' LOTS and HOTS performance (proportion of points earned) for each ELA standard (e.g., Student 1 may have 1.00/1.00 points proportion earned for LOTS and 0.50/1.00 points proportion earned for HOTS for Standard 1), giving us a student-standard observation. We then produce a scatterplot, Appendix Figure 21, documenting 10th grade students' LOTS and HOTS performance for each standard. The y-axis represents the HOTS point proportion earned for each standard, and the x-axis represents the LOTS point

---

based on the absolute fit criteria. While caution is certainly warranted in interpreting any results, the alternative fit index measurement invariance approach detailed in the main text remains preferred

[17] Massachusetts provides a tool, the "Standards Navigator," that maps out which standards connect to one another for MCAS 2.0. Most ELA standards do not connect to other ELA standards in the same grade. The primary exception is Writing standards, which build on both grammar and vocabulary standards. When Massachusetts reports out the standards used for each item in the assessment, all pre-requisite standards are listed—a writing item includes the grammar and vocabulary standards.

[18] It would be concerning to draw conclusions that the HOTS items are completely independent from LOTS items based on standards with almost no items representing that standard.

proportion received. Each bin shows the percentage of 10th grade students that earned each combination of LOTS and HOTS points. For example in the top right bin, 31.8% of student-standard observations received all LOTS points and also all HOTS points; however in the top left bin, only 2.2% of students received almost no LOTS points but nearly all HOTS points. These results suggest that while it is possible for students to answer HOTS items correctly without answering same standard LOTS items correctly, students performing well on LOTS items are far more likely to perform well on same standard HOTS items, supporting the sequential hypothesis.

Discussion

Altogether, our results indicate that leveraging extant statewide data to capture higher-order thinking skills is challenging, but not impossible. Statewide assessments were built to measure students' ELA ability, and the body of validity evidence largely supports the use of these data to capture a unidimensional construct. The fit statistics of the unidimensional model clearly meet Hu's and Bentler's good fit criteria regardless of the test administration. Furthermore, multidimensional measurement models often reveal properties that hint toward unidimensionality—under MCAS 2.0, bidimensional models show nearly perfect correlations between HOTS and LOTS factors, and testlet models leave almost no lingering HOTS or LOTS variation. Even pre-Common Core MCAS 1.0 assessments are not always a great source of information for distinguishing HOTS and LOTS. As a result, we caution researchers and policymakers from defaulting to this approach when attempting to differentiate HOTS from LOTS and learn about HOTS development. Prior work that relies on a similar approach could be revisited to ensure adequate measurement properties in these contexts.

That said, we do find that some opportunities to capture and differentiate HOTS from LOTS exist for certain grades and years using the pre-Common Core exams. The bidimensional model with MCAS 1.0 data features somewhat lower HOTS and LOTS factor correlations, similar fit statistics to an already well-fit unidimensional model, desirable factor loading magnitudes, and the ability to predict postsecondary outcomes even after accounting for LOTS. While many grade-year combinations and 3rd grade assessments are not a good source of information on HOTS, our results suggest that extant data can sometimes capture higher-order thinking skill signal beyond ELA ability in Pre-Common Core aligned assessments. Therefore, these data could be a useful source of information for learning about higher-order skill development relative to lower-order skills. However, the utility of the subscale approach therefore depends on the assessment context and testing regime, warranting careful examination of measurement properties prior to its use.

Although our results suggest researchers should not use Massachusetts' Common Core-aligned assessments to try and differentiate HOTS from LOTS, this does not mean that the approach of coding items for HOTS cannot be used in present-day contexts. As documented in Appendix Table 1 and Table 3, many states' contemporary ELA assessments show similar percentage of HOTS items as Massachusetts during the MCAS 1.0 testing era (2014 and earlier). Instead, researchers interested in this approach may benefit from either examining blueprints when available or coding a subset of assessments before fully committing to this approach to ensure the appropriateness of its use in their context. Additionally, when using exams that are not aligned with Common Core assessments, researchers should examine whether the measurement properties of each specific grade-year assessment warrant the use of that assessment for the purpose of isolating HOTS before proceeding with their analyses.

The good news is that states interested in measuring and advancing student's higher-order thinking may already be well-positioned to do so. Earlier guidelines recommending that 50% of items should measure HOTS for the assessment as a whole to accurately capture HOTS may have been too high. Based on our findings for the Common Core-aligned MCAS 2.0 assessments, the assessment may sufficiently capture higher-order thinking at a threshold closer to ~30% of items measuring HOTS. States consistently and clearly exceeding this HOTS item threshold in their existing assessments could likely use existing assessments to adequately capture HOTS. Other states that either fall closer or below MCAS 1.0's ~20% HOTS item threshold could leverage item-level data to differentiate between HOTS and LOTS to measure and advance students' development of those skills. In future assessment development, these states may also not be far from increasing the number of HOTS items to capture higher-order thinking with the assessment as a whole.

To be clear, our results do not suggest that states and assessment developers should adopt exams more like MCAS 1.0 than MCAS 2.0 just because the Common Core-aligned exams may make it more difficult to distinguish HOTS from LOTS. In fact, from a policy perspective, it may be valuable to increase the share of items on high-stakes exams that capture higher-order skills in order to incentivize their development. Our focus here was really on the question of whether researchers and data analysts can use extant data to retrospectively study higher-order skills and to distinguish them from lower-order skills.

Although this paper has emphasized the importance of higher-order thinking skills, our results also emphasize that the development of students' lower-order thinking skills appears critical for their long-term success. Lower-order thinking skills served as a major predictor of educational attainment outcomes in our earlier findings. Similarly, our findings also suggest that students often must master lower-order thinking skills before demonstrating higher-order thinking skills. In the quest to develop students' higher-order thinking, the measurement and development of lower-order skills should not be abandoned. Although the names of these concepts—"higher" and "lower" order—may imply that one is more important than the other, positioning HOTS as superior to LOTS may be misguided. LOTS should instead be conceptualized as valuable foundational skills. States should continue to assess and cultivate lower-order skills, setting students up not only for later higher-order thinking development but also likely benefiting them in other important outcomes, like educational attainment.

Our results also contribute to enduring scholarly debates regarding the extent to which higher-order skills are domain-specific or domain-general. In other words, whether HOTS in ELA translate to HOTS math and science. We find that in cases where a multidimensional measurement model is defensible, we find that the bidimensional models tend to have more favorable measurement properties than the testlet models. These testlet models are constructed in a way that is more aligned with a domain-general conception of higher-order skills than the bidimensional models. Therefore, although further research is needed, our findings suggest that the development of these skills may require specific strategies tailored to academic subject and that high levels of HOTS in one subject may not necessarily translate to higher HOTS in other domains.

Altogether, given the importance of higher-order thinking skills and the fundamental challenges that have limited other scholars from studying the topic at scale, this research provides novel evidence that flagging individual HOTS assessment items and pairing that coding with item-level response data to capture higher-order skills at scale is possible in some cases, albeit often not recommended. As a consequence, researchers, policymakers, and practitioners

should first examine whether their pre-existing assessments feature enough higher-order thinking skill items to be used to capture HOTS. If not, they can code assessment items and identify assessments with evidence supporting their use for the purpose of isolating HOTS. The use of large language models to automate the coding process could likely make this decision process even more efficient going forward. Still, future research could explore the other subjects (e.g., science), longitudinal measures of higher-order skills (rather than those displayed at a single grade-year), other state assessment contexts, leverage student-level outcomes rather than school-level outcomes, and explore the role of lower-order skills more in-depth. But, as our results hint, these other research directions may not always be fruitful. Nonetheless, this research provides a crucial step in helping us understand an important set of skills in our educational system.

References

Alexander, P. A. (2023). “Here Be Dragons!” Mapping the Realm of Higher-Order, Critical, and Critical-Analytic Thinking. *Educational Psychology Review*, *35*(2), 42.

Ali, U. S., Chang, H.-H., & Anderson, C. J. (2015). Location indices for ordinal polytomous items based on item response theory: IRT location indices for ordinal polytomous items. *ETS Research Report Series*, *2015*(2), 1–13.

Autor, D. H., & Handel, M. J. (2013). Putting Tasks to the Test: Human Capital, Job Tasks, and Wages. *Journal of Labor Economics*, *31*(S1), S59–S96.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, *118*(4), 1279–1333.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Handbook I: cognitive domain. *New York: David McKay*. https://scholar.archive.org/work/l7a35bfkqjge3ictjyt4cb2fsi/access/wayback/https://www.uky.edu/~rsand1/china2018/texts/Bloom%20et%20al%20-Taxonomy%20of%20Educational%20Objectives.pdf

Campbell, D. E. (2006). *What is education’s impact on civic and social engagement*. Proceedings of the OECD Copenhagen Symposium on Social Outcomes of Learning.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255.

Chingos, M. M. (2013). Standardized testing and the common core standards. *Brookings Institution*. https://www.brookings.edu/wp-content/uploads/2016/06/Standardized-Testing-and-the-Common-Core-Standards_FINAL_PRINT.pdf

Christopherson, S. C., & Webb, N. L. (2018). *Alignment analysis of the ACT and SAT with the Georgia standards of excellence for American literature and composition, algebra I, geometry, and biology*. gadoe.org. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/General%20Presentations/Independent_Alignment_Study_ACT_SAT.pdf

Cohodes, S. R. (2016). Teaching to the student: Charter school effectiveness in spite of perverse incentives. *Education Finance and Policy*. https://www.mitpressjournals.org/doi/abs/10.1162/EDFP_a_00175

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., Bennett, R., Gordon, E., Haertel, E., Hakuta, K., & Others. (2013). Criteria for high-quality

assessment. *Stanford Center for Opportunity Policy in Education*, *2*, 171–192.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*. https://academic.oup.com/qje/article-abstract/132/4/1593/3861633

Deming, D. J. (2022). Four facts about human capital. *The Journal of Economic Perspectives: A Journal of the American Economic Association*. https://www.aeaweb.org/articles?id=10.1257%2Fjep.36.3.75

Deming, D., & Kahn, L. B. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*. https://www.journals.uchicago.edu/doi/abs/10.1086/694106

Doorey, N., & Polikoff, M. (2016). Evaluating the content and quality of next generation assessments. *Thomas B. Fordham Institute*. http://files.eric.ed.gov/fulltext/ED565742.pdf

Esdale, R. W. (2021). *An Analysis of Higher-Order Thinking Requirement of the 2018 PSAT/NMSQT* [search.proquest.com]. https://scholarship.shu.edu/cgi/viewcontent.cgi?article=4051&context=dissertations

Facione, P. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report)*. https://philpapers.org/archive/faccta.pdf

Facione, P. A. (1990). *The California Critical Thinking Skills Test--College Level. Technical Report# 2. Factors Predictive of CT Skills*. ERIC. https://eric.ed.gov/?id=ED327550

Gareis, C. R., McMillan, J. H., Smucker, A., & Huang, K. (2021). MAP Growth validation study: An evaluation of the alignment of selected MAP Growth assessments to the Virginia Standards of learning and an exploration of the utility of MAP Growth reports for determining student performance relative to grade level. *Online Submission*. http://files.eric.ed.gov/fulltext/ED618690.pdf

Gilens, M., & Page, B. I. (2014). Testing theories of American politics: Elites, interest groups, and average citizens. *Perspectives on Politics*, *12*(3), 564–581.

Gutmann, A. (1999). *Democratic Education: Revised edition*. Princeton University Press. https://doi.org/10.1515/9781400822911

Herman, J. L., La Torre, D., Epstein, S., & Wang, J. (2016). *Benchmarks for Deeper Learning on Next Generation Tests: A Study of PISA*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). https://eric.ed.gov/?id=ED571800

Hess, Carlock, Jones, & Walkup. (2009). What exactly do “fewer, clearer, and higher standards” really look like in the classroom? Using a cognitive rigor matrix to analyze curriculum, plan lessons, and …. *Retrieved May*.

https://ccsso.confex.com/ccsso/2010/webprogram/Handout/Session1381/cognitive%20rigor%20paper_9%2030%2009%20_2_.pdf

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Iatarola, P., Conger, D., & Long, M. C. (2011). Determinants of high schools' advanced course offerings. *Educational Evaluation and Policy Analysis*, *33*(3), 340–359.

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5), 761–796.

Jochim, A., & McGuinn, P. (2016). The politics of the Common Core assessments: why states are quitting the PARCC and Smarter Balanced testing consortia. *Education Next*, *16*, 44+.

Koenig, J. A. (2011). *Assessing 21st century skills: Summary of a workshop*. National Academies Press. https://books.google.com/books?hl=en&lr=&id=NOwHGOkd0XcC&oi=fnd&pg=PR1&dq=Assessing+21st+century+skills:+Summary+of+a+workshop&ots=zjQhcsdIij&sig=jwtsA4nCbIe--2yYCyywxusHkFk

Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). The perceived effects of the Maryland school performance assessment program. *Retrieved February*. https://www.academia.edu/download/34386607/TECH409.pdf

Koretz, Daniel, Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement Issues and Practice*, *13*(3), 5–16.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3–21.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series*, *2014*(1), 1–23.

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting Diagnostic Classification Models to Responses From IRT-Based Assessment Forms. *Educational and Psychological Measurement*, *78*(3), 357–383.

Liu, Y., & Grusky, D. B. (2013). The Payoff to Skill in the Third Industrial Revolution. *The American Journal of Sociology*, *118*(5), 1330–1374.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics

achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67.

Martone, A., & Sireci, S. G. (2009). Evaluating Alignment Between Curriculum, Assessment, and Instruction. *Review of Educational Research*, *79*(4), 1332–1361.

Mitani, H. (2021). Test Score Gaps in Higher Order Thinking Skills: Exploring Instructional Practices to Improve the Skills and Narrow the Gaps. In *AERA Open* (Vol. 7, p. 233285842110164). https://doi.org/10.1177/23328584211016470

Papay, J. P. (2011). Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Educational Research Journal*, *48*(1), 163–193.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. Developmental Review: DR, 41, 71–90.

Peterson, P. E., Barrows, S., & Gift, T. (2016). After Common Core, states set rigorous standards. *Education Next*, *16*(3), 9–15.

Philpot, R., Lindquist, M., Mullis, I. V. S., & Aldrich, C. E. A. (2021). *TIMSS 2023 Mathematics Framework*. ERIC.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model: Formal relations among IRT multidimensional models. *Journal of Educational Measurement*, *47*(3), 361–372.

Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, *25*(1), 6–14.

Schleicher, A. (2024). *Quality standards for PISA* (EDU/PISA/GB(2024)9). Organisation for Economic Co-operation and Development.

Schueler, B. & Larned, K. (2023). Interscholastic policy debate promotes critical thinking and college-going: Evidence from Boston Public Schools. *Educational Evaluation and Policy Analysis*, 47(1).

Stitzlein, S. M. (2021). Defining and Implementing Civic Reasoning and Discourse: Philosophical and Moral Foundations for Research and Practice. In L. C. Ed., W. G. Ed., & D. D. Ed. (Eds.), *Educating for Civic Reasoning and Discourse* (pp. 23–52).

Webb. (2002). Depth-of-knowledge levels for four content areas. *Language Arts*. http://ossucurr.pbworks.com/w/file/fetch/49691156/Norm%20web%20dok%20by%20subject%20area.pdf

Wixson, K. K., Fisk, M. C., Dutro, E., & McDaniel, J. (2002). *The Alignment of State Standards and Assessments in Elementary Reading*. Center for the Improvement of Early Reading Achievement. https://eric.ed.gov/?id=ED474625

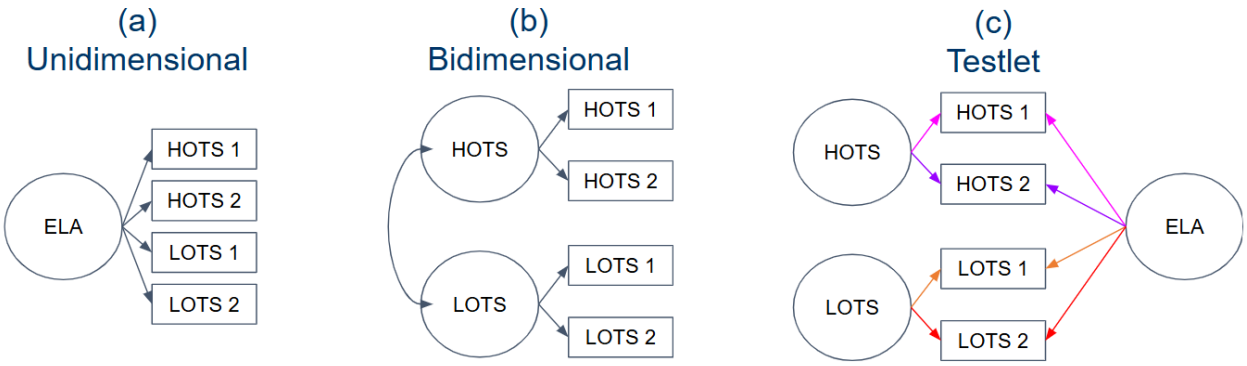
Yuan, K., & Le, V.-N. (2012). *Estimating the Percentage of Students Who Were Tested on Cognitively Demanding Items Through the State Achievement Tests*. RAND Corporation. https://doi.org/10.7249/wr967

Yuan, K., & Le, V.-N. (2014). Measuring Deeper Learning through Cognitively Demanding Test Items: Results from the Analysis of Six National and International Exams. Research Report. *RAND Corporation*. http://www.rand.org/pubs/research_reports/RR483.html

Yuan, Y., & Engelhard, G., Jr. (2023). Using linear logistic Rasch models to examine cognitive complexity and linguistic cohesion in science items. In *Contemporary Trends and Issues in Science Education* (pp. 455–482). Springer International Publishing.

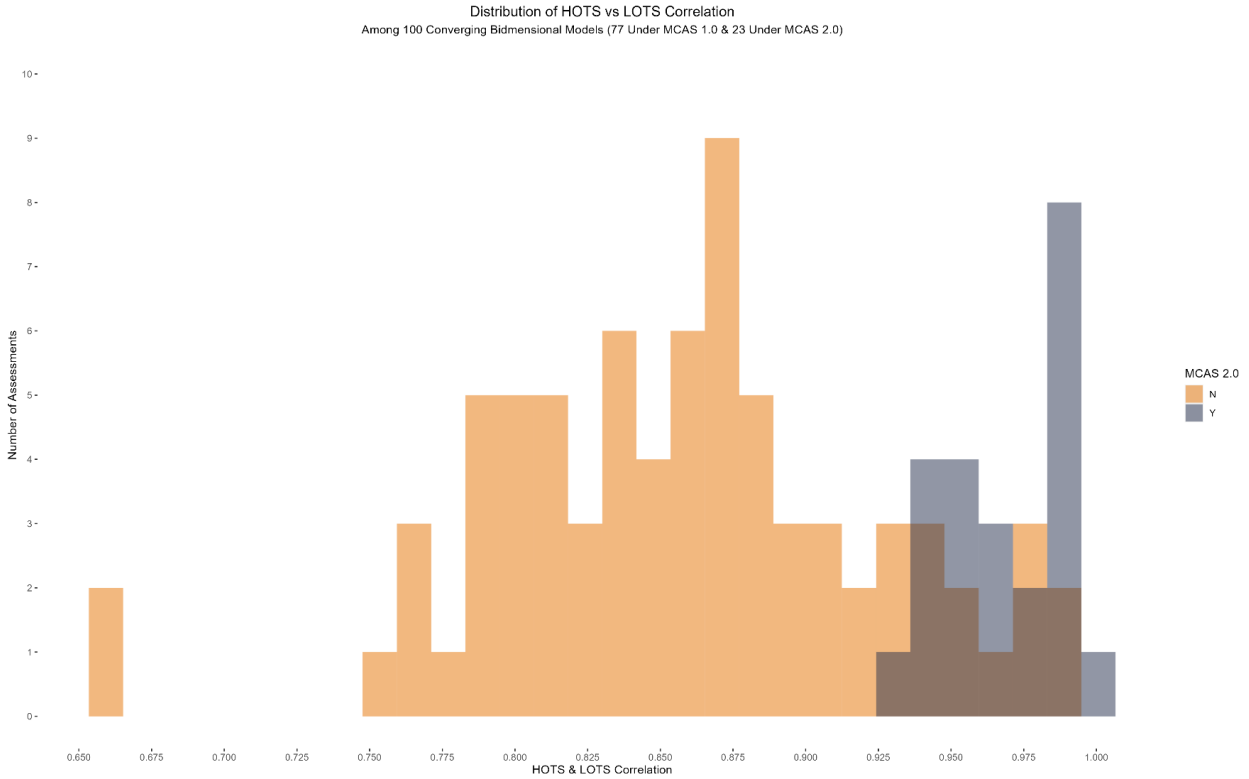
Zahner, D., & James, J. K. (2015). Predictive Validity of a Critical Thinking Assessment for Post-College Outcomes. Council for Aid to Education. https://eric.ed.gov/?id=ED582251_

Figure 1: Measurement Models



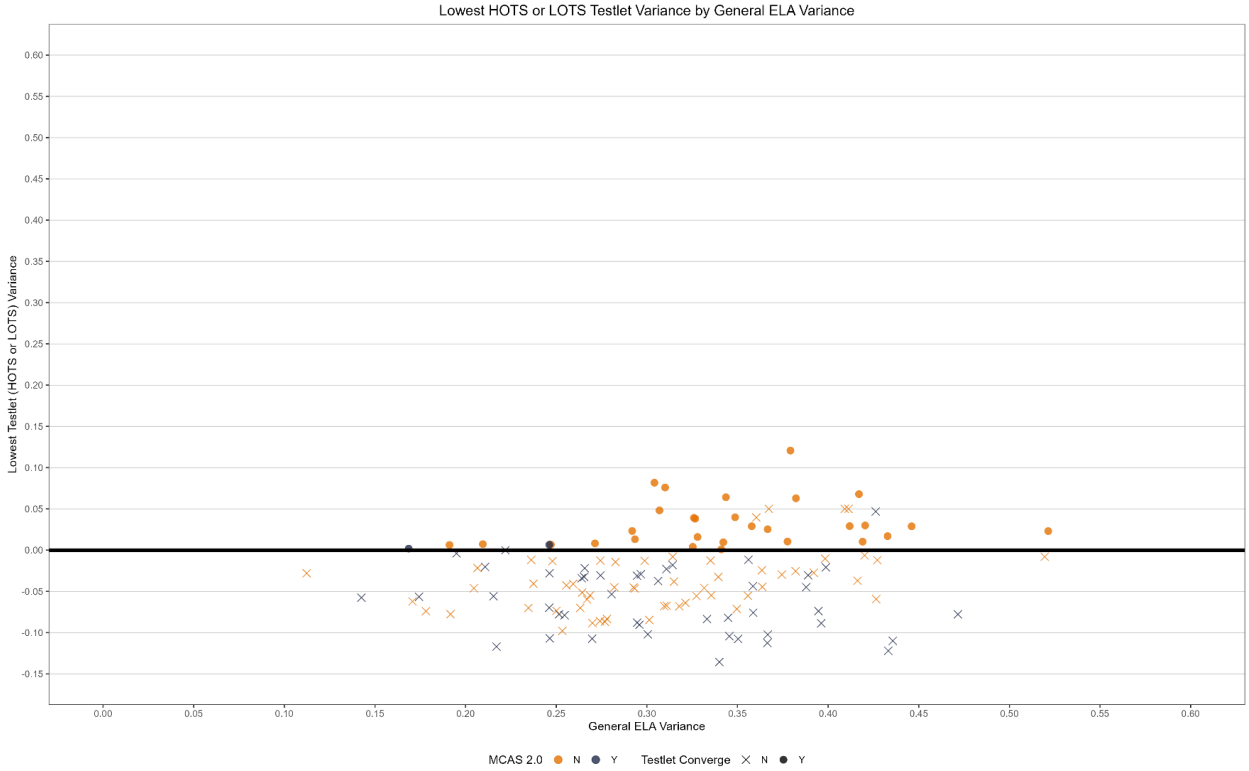
Note: This figure provides a visual representation of the (a) unidimensional, (b) bidimensional, and (c) testlet measurement models. The circles (e.g., ELA) are the latent variable or factor, and the single-headed arrows represent the relationship of the latent variable onto manifest variables (higher-order thinking or lower-order thinking skill items). For the bidimensional model, the double-headed arrows show that HOTS and LOTS factors vary with one another. For the testlet model, the matched colored arrows indicate that the loadings of the specific factors (HOTS and LOTS) and the general factor (ELA) are identical to each other.

Figure 2: Distribution of Bidimensional Higher-Order and Lower-Order Thinking Skill Correlation by Testing Regime



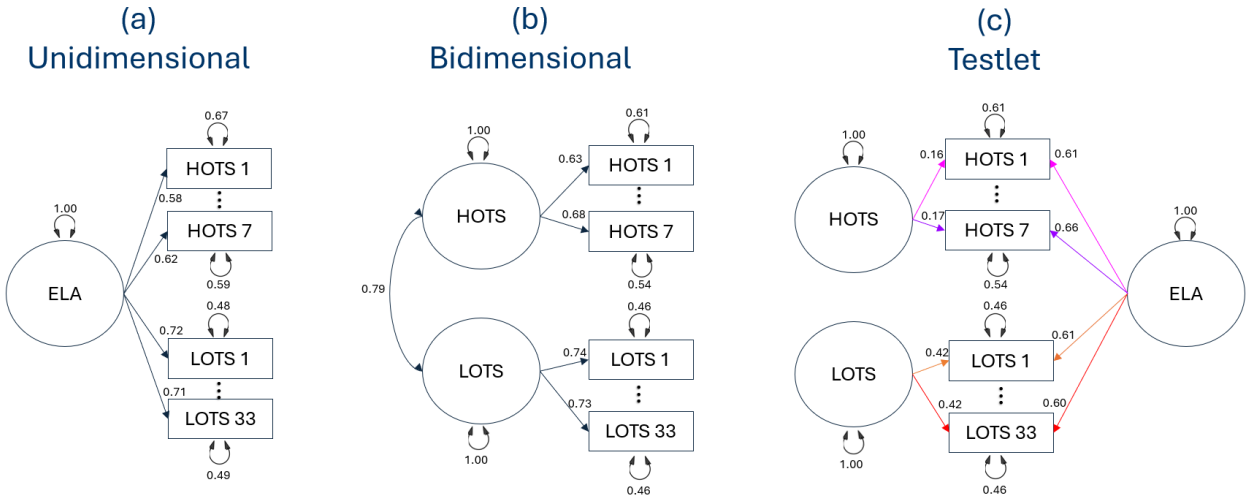
Note: This figure provides a visual representation of the HOTS and LOTS factor correlations from the bidimensional measurement model. In particular, it specifies the correlation on the x-axis along with the number of assessments that fall within that correlation range on the y-axis. Each test regime, either Pre-Common Core Aligned MCAS 1.0 or Post-Common Core Aligned MCAS 2.0, are separately displayed.

Figure 3: Lowest HOTS or LOTS Testlet Variance by General ELA Variance Scatterplot



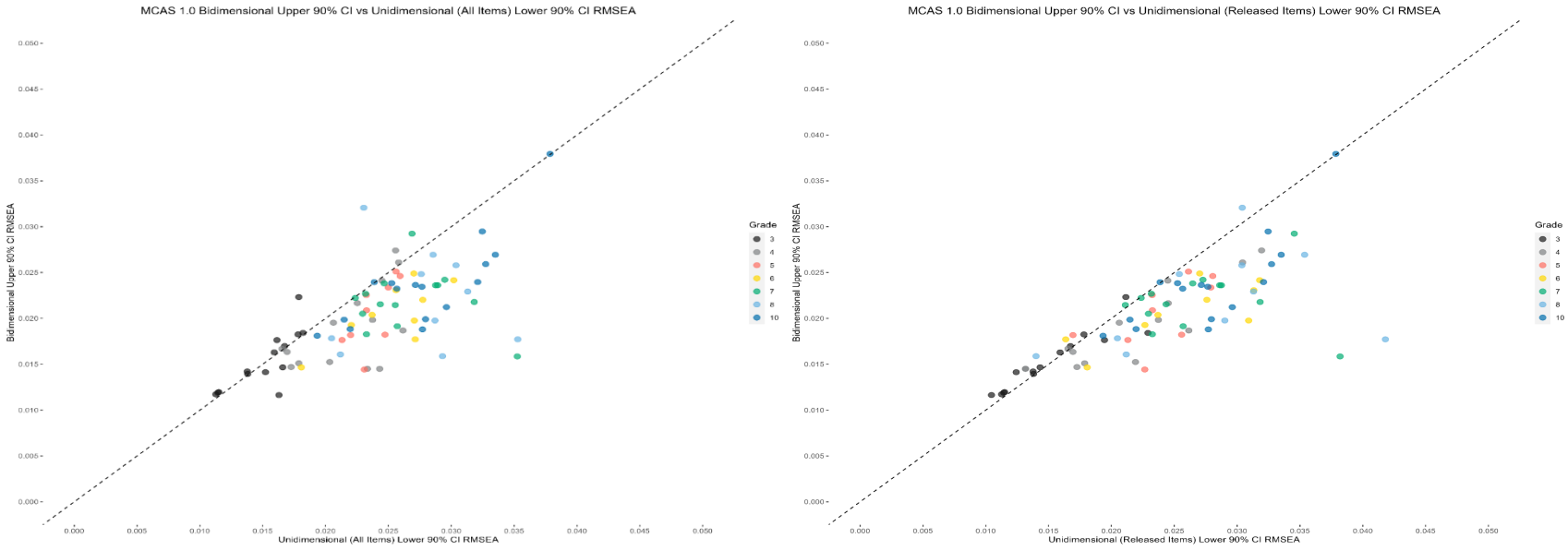
Note: This figure provides a visual representation of the testlet model's variance. Specifically, the x-axis displays the variance of the General ELA factor, and the y-axis shows the variance of either the LOTS or HOTS factor (whichever is lowest). The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged.

Figure 4: 10th Graders in 2014 Measurement Model Path Diagrams



Note: This figure provides an abbreviated visual representation of the parameters of the (a) unidimensional, (b) bidimensional, and (c) testlet measurement models. The full parameters can be found in Table 5. The circles (e.g., ELA) are the latent variable or factor, and the single-headed arrows, and corresponding values, represent the relationship of the latent variable onto manifest variables (higher-order thinking or lower-order thinking skill items). The double-headed arrows that point back to a factor reflect a value of 1 to indicate that the model is standardized, and the double-headed arrows that point back to an item refers to the residual variance. For the bidimensional model, the double-headed arrows show that HOTS and LOTS factors vary with one another. For the testlet model, the matched colored arrows indicate that the loadings of the specific factors (HOTS and LOTS) and the general factor (ELA) are identical to each other.

Figure 5: MCAS 1.0 Bidimensional Measurement Model Scatterplot Comparison Results Using RMSEA Confidence Intervals



Note: These figures compare the bidimensional and unidimensional (version with All items and only Released items) models' RMSEA with confidence intervals for each MCAS 1.0 grade. The bidimensional models use the upper 90% confidence interval RMSEA, while the unidimensional models use the lower 90% confidence interval RMSEA, providing a conservative comparison between the models. If the point falls above the 45-degree line, then the unidimensional RMSEA statistic is better (lower) than the bidimensional RMSEA statistic, suggesting that the unidimensional model is better. The reverse would suggest that the bidimensional fit statistic is superior.

Table 1: Demographics

Demographic	All Years	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2017	2018	2019	2021	2022	2023
Asian	**0.06**	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.07	0.07	0.07	0.07	0.07	0.07
Black	**0.09**	0.08	0.09	0.09	0.09	0.09	0.08	0.08	0.08	0.08	0.08	0.08	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09
Hispanic	**0.16**	0.10	0.11	0.11	0.11	0.12	0.13	0.13	0.14	0.14	0.15	0.15	0.16	0.16	0.17	0.19	0.20	0.21	0.22	0.24	0.25
White	**0.66**	0.75	0.75	0.75	0.74	0.74	0.72	0.72	0.71	0.70	0.69	0.68	0.67	0.66	0.65	0.61	0.60	0.59	0.57	0.55	0.54
Economically Disadvantaged	**0.35**	0.24	0.27	0.28	0.28	0.29	0.30	0.30	0.31	0.32	0.35	0.36	0.37	0.38	0.39	0.34	0.35	0.36	0.39	0.47	0.45
Male	**0.51**	0.50	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
Female	**0.49**	0.48	0.48	0.48	0.49	0.48	0.48	0.48	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.48	0.48
Special Education	**0.18**	0.16	0.16	0.17	0.16	0.17	0.17	0.18	0.18	0.18	0.18	0.19	0.18	0.18	0.18	0.19	0.18	0.20	0.19	0.21	0.21
English Language Learner	**0.07**	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.07	0.07	0.07	0.07	0.09	0.09	0.10	0.09	0.11	0.12
Observations	**9,169,761**	294,952	301,845	303,094	297,350	297,896	517,776	513,423	553,938	553,911	561,090	561,412	560,712	560,316	560,431	507,552	495,834	505,661	301,161	483,511	437,896

Note: Values in each row (except years and observations) represent fractions (e.g., across all years, 0.06, or 6%, of Massachusetts students were Asian). The definition of economically disadvantaged has changed over time in Massachusetts. The original definition from 2001 to 2014 consisted of whether a student was eligible for free-reduced price lunch. In the 2011-2012 school year, the U.S. Department of Agriculture introduced the Community Eligible Program, which allowed all students in schools/districts with high concentrations of low income students to receive free meals—as a consequence, data on free-reduced lunch eligibility became less available as more schools and districts adopted the Community Eligible Program. In the 2014-2015 school year, the state shifted to "economically disadvantaged," which flags a student as economically disadvantaged if the student participates in the Supplemental Nutrition Program, the Transitional Assistance for Families with Dependent Children, the Department of Children and Families' foster care program, or Medicaid. Finally, on top of the previously mentioned components of economically disadvantaged, starting in the 2021-2022 school year, Massachusetts further added a supplemental data collection form to identify students that meet the 185% of federal poverty level threshold but were not identified under the previous criteria. Massachusetts also included students that were reported by the district as "homeless" as part of their economically disadvantaged definition.

Table 2: Higher-Order Thinking Skills Coding Breakdown

	DOK Coding Results	All Years	All Years Except 2015 or 2016	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2021	2022	2023
Overall	Total Items	5169	4607	206	162	162	162	162	282	282	282	282	281	281	281	281	281	281	281	190	201	176	218	217	218
	Total Released Items	3570	3286	202	162	162	162	162	282	282	282	139	140	136	147	151	138	146	138	101	121	95	137	147	138
	Number of HOTS Items	668	625	23	18	24	28	26	54	47	54	24	27	21	26	23	24	22	21	21	35	33	35	41	41
	HOTS Percentage	19%	19%	11%	11%	15%	17%	16%	19%	17%	19%	17%	19%	15%	18%	15%	17%	15%	15%	21%	29%	35%	26%	28%	30%
3rd	Total Released Items	564	524	42	42	42	42	42	42	42	42	13	18	17	17	18	18	23	17	12	9	8	18	20	20
	Number of HOTS Items	52	47	2	2	2	2	3	3	4	4	1	3	3	2	2	2	3	2	1	2	2	3	2	2
	HOTS Percentage	9%	9%	5%	5%	5%	5%	7%	7%	10%	10%	8%	17%	18%	12%	11%	11%	13%	12%	8%	22%	25%	17%	10%	10%
4th	Total Released Items	552	517	40	40	40	40	40	40	40	40	17	16	16	17	18	17	17	18	12	18	8	18	20	20
	Number of HOTS Items	88	83	5	5	6	10	5	7	7	6	4	1	1	2	3	3	2	3	2	5	2	1	3	5
	HOTS Percentage	16%	16%	13%	13%	15%	25%	13%	18%	18%	15%	24%	6%	6%	12%	17%	18%	12%	17%	17%	28%	25%	6%	15%	25%
5th	Total Released Items	346	312						40	40	40	17	16	16	18	18	17	17	17	7	9	17	19	19	19
	Number of HOTS Items	71	65						11	7	6	2	4	1	3	2	3	3	3	1	4	8	3	5	5
	HOTS Percentage	21%	21%						28%	18%	15%	12%	25%	6%	17%	11%	18%	18%	18%	14%	44%	47%	16%	26%	26%
6th	Total Released Items	352	319						40	40	40	18	17	17	18	24	16	17	16	10	17	8	20	19	15
	Number of HOTS Items	72	66						6	6	10	4	2	3	3	4	1	5	1	1	5	3	5	6	7
	HOTS Percentage	20%	21%						15%	15%	25%	22%	12%	18%	17%	17%	6%	29%	6%	10%	29%	38%	25%	32%	47%
7th	Total Released Items	542	513	40	40	40	40	40	40	40	40	17	16	18	19	14	18	16	13	10	10	16	20	20	15
	Number of HOTS Items	108	103	5	5	8	7	6	6	8	7	2	3	4	6	4	3	3	2	1	4	7	8	6	3
	HOTS Percentage	20%	20%	13%	13%	20%	18%	15%	15%	20%	18%	12%	19%	22%	32%	29%	17%	19%	15%	10%	40%	44%	40%	30%	20%
8th	Total Released Items	374	341	40					40	40	40	17	17	12	18	19	12	16	17	10	18	8	12	19	19
	Number of HOTS Items	78	73	5					7	6	9	2	4	1	4	2	5	2	3	2	7	3	2	7	7
	HOTS Percentage	21%	21%	13%					18%	15%	23%	12%	24%	8%	22%	11%	42%	13%	18%	20%	39%	38%	17%	37%	37%
10th	Total Released Items	840	760	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	30	30	30	30
	Number of HOTS Items	199	188	6	6	8	9	12	14	9	12	9	10	8	6	6	7	4	7	13	8	8	13	12	12
	HOTS Percentage	24%	25%	15%	15%	20%	23%	30%	35%	23%	30%	23%	25%	20%	15%	15%	18%	10%	18%	33%	20%	27%	43%	40%	40%

Note: Though Massachusetts generally uses ~40 items per assessment (~30 in 2017 onward), the state releases only a fraction of the assessment items. All released assessment items were coded using Webb's Depth of Knowledge (DOK). Items coded as DOK 1 (Recall/Reproduction) or 2 (Skills/Concept) were classified as lower-order thinking skills; items coded as DOK 3 (Strategic Thinking) or 4 (Extended Thinking) were classified as higher-order thinking skills (HOTS). The results of the coding across all years and grades as well as each individual year are reported in this table. Additionally, since the public Massachusetts student-level performance data do not include the 2014-2015 and 2015-2016 school years, the results across all but those years are reported as well.

Table 3: Higher-Order Thinking Skills Item Coding Breakdown by Test Regime

Item Coding by Test Regime	Total Items	Total Released Items	Number HOTS Items	HOTS Percentage	HOTS Points Percentage
All Grades & Years	4607	3286	625	19%	40%
MCAS 1.0	3467	2627	440	17%	35%
3rd	583	437	35	8%	20%
4th	561	421	65	15%	35%
5th	360	222	39	18%	37%
6th	360	230	39	17%	37%
7th	561	422	74	18%	37%
8th	401	255	45	18%	38%
10th	641	640	143	22%	40%
MCAS 2.0	1140	659	185	28%	53%
3rd	193	87	12	14%	37%
4th	194	96	18	19%	42%
5th	194	90	26	29%	53%
6th	193	89	27	30%	57%
7th	195	91	29	32%	58%
8th	194	86	28	33%	59%
10th	121	120	45	38%	60%

Note: Table 3 reorganizes the results from Table 2 to highlight the key distinctions that emerge by testing regime. Pre-Common Core Aligned MCAS 1.0 assessments and Post-Common Core Aligned MCAS 2.0 assessments feature different degrees of HOTS items and percentage of the overall assessment.

Table 4: Unidimensional Measurement Model Fit Statistics by Testing Regime

Assessment Combinations		Unidimensional (All Items) *Average* Fit Statistics					Unidimensional (Released Items) *Average* Fit Statistics				
Group	Converged Models (N)	Met H&B (1999) Good Fit	RMSEA	SRMR	TLI	CFI	Met H&B (1999) Good Fit	RMSEA	SRMR	TLI	CFI
All Grades & Years	133	100%	0.024	0.031	0.987	0.988	100%	0.022	0.029	0.989	0.990
MCAS 1.0	86	100%	0.024	0.032	0.985	0.986	100%	0.024	0.032	0.985	0.987
3rd	14	100%	0.015	0.025	0.994	0.995	100%	0.016	0.025	0.994	0.995
4th	14	100%	0.022	0.030	0.987	0.988	100%	0.022	0.029	0.988	0.989
5th	9	100%	0.024	0.030	0.987	0.987	100%	0.024	0.030	0.987	0.988
6th	9	100%	0.026	0.032	0.985	0.985	100%	0.026	0.032	0.985	0.987
7th	14	100%	0.027	0.035	0.983	0.984	100%	0.027	0.034	0.982	0.983
8th	10	100%	0.028	0.035	0.983	0.984	100%	0.028	0.034	0.983	0.985
10th	16	100%	0.028	0.038	0.979	0.980	100%	0.028	0.038	0.979	0.980
MCAS 2.0	47	100%	0.024	0.029	0.991	0.992	100%	0.017	0.023	0.995	0.996
3rd	7	100%	0.018	0.022	0.994	0.995	100%	0.013	0.018	0.997	0.998
4th	7	100%	0.021	0.028	0.992	0.992	100%	0.016	0.025	0.995	0.996
5th	7	100%	0.023	0.030	0.991	0.992	100%	0.017	0.024	0.994	0.996
6th	7	100%	0.028	0.029	0.990	0.991	100%	0.017	0.021	0.996	0.996
7th	8	100%	0.026	0.029	0.990	0.991	100%	0.018	0.022	0.995	0.996
8th	7	100%	0.027	0.034	0.989	0.990	100%	0.016	0.024	0.995	0.996
10th	4	100%	0.027	0.028	0.989	0.990	100%	0.027	0.028	0.989	0.990

Note: The table shows the number of converged unidimensional models (both versions with All items or only Released items) and percentage that met Hu and Bentler (1999) 'good fit' criteria as well as the average Root Mean Squared Error of Approximation (RMSEA), Standardized Root Mean Squared Residuals (SRMR), Comparative Fit Index (CFI), and the Tucker-Lewis Index (TLI) across all grades and years or for each testing regime and grade within testing regime.

Table 5: Full Model Examples by Testing Regime

10th Graders in 2014 (MCAS 1.0)									
Model	Unidimensional		Bidimensional			Testlet			
			HOTS	LOTS		General	HOTS	LOTS	
Unstandardized Variance	0.52		0.39	0.54		0.37	0.03	0.18	
Correlation	N/A		0.79			0.00			
Item	Loadings	Residual Varainces	Loadings		Residual Variances	Loadings			Residual Variances
HOTS 1	0.58	0.67	0.63		0.61	0.61	0.16		0.61
HOTS 2	0.69	0.53	0.75		0.44	0.72	0.19		0.44
HOTS 3	0.73	0.46	0.80		0.37	0.77	0.20		0.37
HOTS 4	0.55	0.70	0.59		0.65	0.57	0.15		0.65
HOTS 5	0.75	0.82	0.82		0.33	0.79	0.21		0.33
HOTS 6	0.29	0.92	0.31		0.90	0.30	0.08		0.90
HOTS 7	0.62	0.59	0.68		0.54	0.66	0.17		0.54
LOTS 1	0.72	0.48		0.74	0.46	0.61		0.42	0.46
LOTS 2	0.38	0.86		0.39	0.85	0.32		0.22	0.85
LOTS 3	0.47	0.78		0.49	0.76	0.40		0.28	0.76
LOTS 4	0.59	0.66		0.60	0.64	0.49		0.34	0.64
LOTS 5	0.47	0.78		0.48	0.77	0.40		0.28	0.77
LOTS 6	0.46	0.79		0.47	0.78	0.38		0.27	0.78
LOTS 7	0.42	0.82		0.44	0.81	0.36		0.25	0.81
LOTS 8	0.54	0.71		0.56	0.69	0.46		0.32	0.69
LOTS 9	0.58	0.66		0.60	0.64	0.49		0.34	0.64
LOTS 10	0.67	0.56		0.68	0.53	0.56		0.39	0.53
LOTS 11	0.71	0.50		0.73	0.47	0.60		0.41	0.47
LOTS 12	0.57	0.68		0.58	0.66	0.48		0.33	0.66
LOTS 13	0.45	0.80		0.46	0.79	0.38		0.26	0.79
LOTS 14	0.75	0.43		0.77	0.40	0.63		0.44	0.40
LOTS 15	0.58	0.66		0.60	0.64	0.49		0.34	0.64
LOTS 16	0.52	0.73		0.53	0.71	0.44		0.31	0.71
LOTS 17	0.56	0.69		0.58	0.67	0.47		0.33	0.67
LOTS 18	0.62	0.61		0.64	0.59	0.52		0.37	0.59
LOTS 19	0.48	0.77		0.49	0.76	0.40		0.28	0.76
LOTS 20	0.59	0.65		0.60	0.63	0.50		0.35	0.63
LOTS 21	0.46	0.78		0.48	0.77	0.39		0.27	0.77
LOTS 22	0.47	0.78		0.49	0.76	0.40		0.28	0.76
LOTS 23	0.53	0.72		0.54	0.70	0.45		0.31	0.70
LOTS 24	0.60	0.64		0.62	0.62	0.50		0.35	0.62
LOTS 25	0.60	0.64		0.61	0.62	0.50		0.35	0.62
LOTS 26	0.62	0.61		0.64	0.59	0.52		0.37	0.59
LOTS 27	0.44	0.80		0.46	0.79	0.37		0.26	0.79
LOTS 28	0.68	0.54		0.69	0.52	0.57		0.40	0.52
LOTS 29	0.64	0.59		0.65	0.57	0.54		0.37	0.57
LOTS 30	0.42	0.82		0.43	0.81	0.35		0.25	0.81
LOTS 31	0.50	0.75		0.52	0.73	0.42		0.30	0.73
LOTS 32	0.46	0.78		0.48	0.77	0.39		0.27	0.77
LOTS 33	0.71	0.49		0.73	0.46	0.60		0.42	0.46

10th Graders in 2023 (MCAS 2.0)									
Model	Unidimensional		Bidimensional			Testlet			
			HOTS	LOTS		General	HOTS	LOTS	
Unstandardized Variance	0.32		0.23	0.33		0.27	-0.03	0.06	
Correlation	N/A		0.96			0.00			
Item	Loadings	Residual Variances	Loadings		Residual Variances	Loadings			Residual Variances
HOTS 1	0.48	0.77	0.48		0.77	0.52	N/A		0.77
HOTS 2	0.57	0.67	0.58		0.67	0.62	N/A		0.67
HOTS 3	0.53	0.71	0.80		0.36	0.85	N/A		0.36
HOTS 4	0.53	0.72	0.53		0.71	0.57	N/A		0.71
HOTS 5	0.58	0.67	0.58		0.66	0.62	N/A		0.66
HOTS 6	0.78	0.39	0.79		0.38	0.84	N/A		0.38
HOTS 7	0.66	0.56	0.67		0.55	0.71	N/A		0.55
HOTS 8	0.64	0.59	0.64		0.58	0.69	N/A		0.58
HOTS 9	0.60	0.64	0.60		0.64	0.64	N/A		0.64
HOTS 10	0.81	0.35	0.81		0.34	0.87	N/A		0.34
HOTS 11	0.58	0.66	0.58		0.66	0.62	N/A		0.66
HOTS 12	0.69	0.53	0.69		0.52	0.74	N/A		0.52
LOTS 1	0.57	0.68		0.57	0.67	0.52		0.25	0.25
LOTS 2	0.48	0.77		0.49	0.76	0.44		0.22	0.22
LOTS 3	0.56	0.68		0.57	0.67	0.51		0.25	0.25
LOTS 4	0.39	0.85		0.40	0.84	0.36		0.18	0.84
LOTS 5	0.23	0.95		0.24	0.94	0.21		0.10	0.94
LOTS 6	0.67	0.55		0.68	0.54	0.61		0.30	0.54
LOTS 7	0.69	0.53		0.70	0.51	0.62		0.31	0.51
LOTS 8	0.56	0.68		0.57	0.67	0.51		0.25	0.67
LOTS 9	0.63	0.60		0.64	0.59	0.57		0.28	0.59
LOTS 10	0.51	0.74		0.52	0.73	0.46		0.23	0.73
LOTS 11	0.48	0.77		0.48	0.77	0.43		0.21	0.77
LOTS 12	0.48	0.77		0.48	0.77	0.43		0.21	0.77
LOTS 13	0.59	0.65		0.60	0.64	0.54		0.27	0.64
LOTS 14	0.75	0.43		0.76	0.42	0.68		0.34	0.42
LOTS 15	0.67	0.55		0.68	0.54	0.61		0.30	0.54
LOTS 16	0.57	0.68		0.58	0.67	0.52		0.26	0.67
LOTS 17	0.69	0.53		0.70	0.51	0.62		0.31	0.51
LOTS 18	0.60	0.63		0.61	0.62	0.55		0.27	0.62

Note: The table shows the full model for a MCAS 1.0 assessment (10th graders in 2014) and MCAS 2.0 assessment (10th graders in 2023). This includes the unstandardized variance, correlation (when applicable), and standardized loadings and standardized residual variances for each the unidimensional, bidimensional, and testlet measurement models.

Table 6: Bidimensional Measurement Models Factor Correlations

Bidimensional (Released Items)		**HOTS and LOTS Correlations**	**Degree of Discriminant Validity Concern (Rönkkö and Cho, 2022)**			
Assessment Combinations	Converged Models (N)	Average	No Problem {Corr < 0.8} (%)	Marginal Problem {0.8 ≤ Corr < 0.9} (%)	Moderate Problem {0.9 ≤ Corr < 1.0} (%)	Severe Problem {Corr = 1.0} (%)
All Grades & Years	100	0.883	13	45	41	1
MCAS 1.0	77	0.858	17	58	25	0
3rd	10	0.914	20	10	70	0
4th	12	0.861	17	58	25	0
5th	8	0.855	13	75	13	0
6th	8	0.832	13	88	0	0
7th	14	0.846	21	57	21	0
8th	9	0.835	33	44	22	0
10th	16	0.858	6	75	19	0
MCAS 2.0	23	0.968	0	0	96	4
3rd	3	0.990	0	0	100	0
4th	4	0.961	0	0	100	0
5th	3	0.990	0	0	67	33
6th	3	0.967	0	0	100	0
7th	2	0.966	0	0	100	0
8th	4	0.959	0	0	100	0
10th	4	0.953	0	0	100	0

Note: The correlations between HOTS and LOTS are extracted from assessment combinations with bidimensional measurement models that fit the data. These factor correlations are then averaged across all grades and years. This process repeats for any individual year or grade. Then, using the extracted correlations, the number of assessment combinations with bidimensional measurement models that fit the data are categorized into one group of discriminant validity concern before being divided by the total number of assessment combinations to produce a percentage.

Table 7: Bidimensional Measurement Model Comparison Results

Bidimensional (Released Items)	**Assessments** (N)	**Converged Models** (N)	**Met Hu & Bentler (1999) Good Fit** (% of Converged Models)	**Fit > Unidimensional (All Items)** (% of Converged Models)	**Fit > Unidimensional (Released Items)** (% of Converged Models)	**Fit > Unidimensional (Both All & Released)** (% of Converged Models)
MCAS 1.0	86	77	100%	82%	84%	78%
3rd	14	10	100%	70%	70%	60%
4th	14	12	100%	75%	92%	75%
5th	9	8	100%	75%	75%	75%
6th	9	8	100%	75%	88%	75%
7th	14	14	100%	93%	93%	86%
8th	10	9	100%	89%	78%	78%
10th	16	16	100%	88%	88%	88%

Note: Since Massachusetts's public data provides student-level item performance for all items, we are able to estimate a unidimensional measurement model for using all items. In addition, we estimate a version with only released assessment items to create a similar comparison to the bidimensional and testlet models which use only released assessment items. The columns of the table then report out the number of assessment combinations (grade-year-modality/session), the number of models that converge without errors (e.g., negative variances or impossible correlations), and the percentage of models that meet Hu's and Bentler's (1999) 'good fit' criteria. The subsequents columns of the table compare the fit statistics between the bidimensional models that converge to both unidimensional counterparts. The values represent the percentage of models that provide a superior fit across all fit statistics.

Table 8: Bidimensional Measurement Models Factor Loadings

Bidimensional (Released Items)		**Higher-Order Thinking Skills *Average* Standardized Factor Loadings**				**Lower-Order Thinking Skills *Average* Standardized Factor Loadings**			
Assessment Combinations	Converged Models (N)	Minimum	Mean	Standard Deviation	Max	Minimum	Mean	Standard Deviation	Max
MCAS 1.0	77	0.53	0.66	0.09	0.76	0.34	0.58	0.11	0.75
3rd	10	0.57	0.62	0.06	0.67	0.36	0.61	0.11	0.78
4th	12	0.52	0.64	0.08	0.73	0.32	0.57	0.11	0.74
5th	8	0.61	0.69	0.07	0.76	0.38	0.58	0.09	0.72
6th	8	0.56	0.67	0.09	0.76	0.40	0.59	0.10	0.75
7th	14	0.54	0.67	0.10	0.77	0.33	0.57	0.11	0.75
8th	9	0.56	0.69	0.09	0.78	0.38	0.59	0.10	0.76
10th	16	0.46	0.66	0.12	0.80	0.29	0.55	0.12	0.77

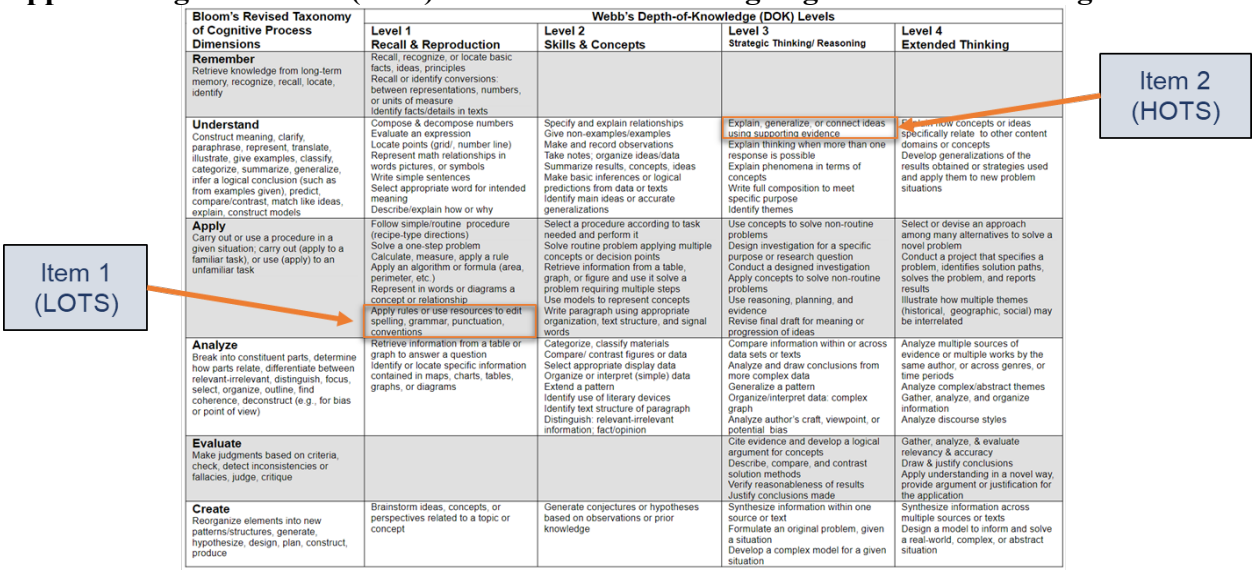
Note: The minimum, mean, standard deviation, and maximum standardized factor loading are identified and extracted for each latent variable in an assessment combination for which the bidimensional measurement model fits the data. To produce the average standardized factor loadings, the minimum standardized factor loading across all assessment combinations with bidimensional models fitting the data are averaged (repeats for mean, standard deviation, and maximum). Each year or grade follows an identical process.

Table 9: Bidimensional Measurement Models Predictive Validity

Bidimensional Models	School's Percentage of Students Graduating High School Within 4-Years			School's Percentage of Students Enrolling in Postsecondary Institutions Within 16 Months		
HOTS	3.80*** (0.342)		0.26 (0.312)	6.26*** (0.390)		1.26** (0.440)
LOTS		4.04*** (0.353)	3.80*** (0.423)		6.54*** (0.396)	5.36*** (0.510)
Observations	888906	888906	888906	886948	886948	886948
Adj R^2	0.154	0.166	0.166	0.176	0.188	0.189
	School's Percentage of Students Enrolling in 2-Year Postsecondary Institutions Within 16 Months			School's Percentage of Students Enrolling in 4-Year Postsecondary Institutions Within 16 Months		
HOTS	-2.34*** (0.196)		-0.19 (0.362)	8.61*** (0.498)		1.46* (0.698)
LOTS		-2.49*** (0.204)	-2.31*** (0.397)		9.02*** (0.501)	7.67*** (0.752)
Observations	886948	886948	886948	886948	886948	886948
Adj R^2	0.104	0.112	0.112	0.164	0.178	0.179

Note: * p <0.05, ** p <0.01, and *** p<0.001. Percentage of school's postsecondary enrollment was created by using dividing the number of postsecondary enrollments within 16 months by the number of high schoolers that originally started in that school, for example, in 2009. These results use the "Unadjusted" populations (i.e., they account for students transferring in or out of the school). All Models include cohort fixed effects. School's Percentage of Students Graduating High School Within 4-Years Mean Across All Years: 86.626%. School's Percentage of Students Enrolling in Postsecondary Institutions Within 16 Months Mean Across All Years: 66.696%. School's Percentage of Students Enrolling in 2-Year Postsecondary Institutions Within 16 Months Mean Across All Years: 18.101%. School's Percentage of Students Enrolling in 4-Year Postsecondary Institutions Within 16 Months Mean Across All Years: 48.587%.

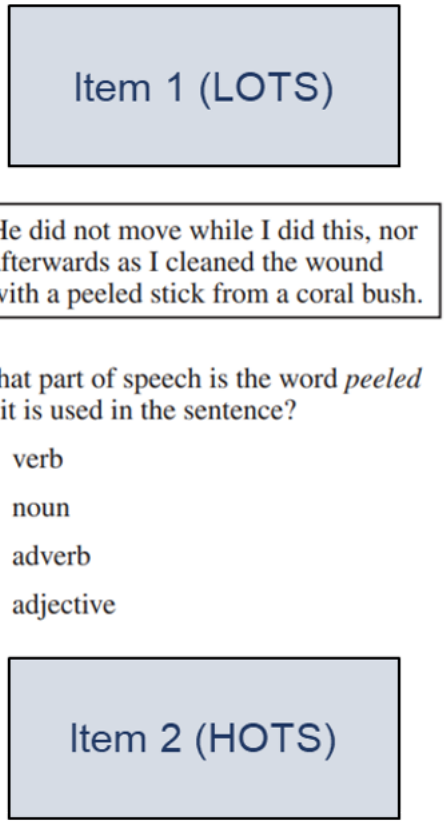
Appendix Figure 1: Hess (2009) Framework for Coding Higher-Order Thinking Skills



Bloom's Revised Taxonomy of Cognitive Process Dimensions	Webb's Depth-of-Knowledge (DOK) Levels			
	Level 1 Recall & Reproduction	**Level 2** Skills & Concepts	**Level 3** Strategic Thinking/ Reasoning	**Level 4** Extended Thinking
Remember Retrieve knowledge from long-term memory, recognize, recall, locate, identify	Recall, recognize, or locate basic facts, ideas, principles Recall or identify conversions: between representations, numbers, or units of measure Identify facts/details in texts			
Understand Construct meaning, clarify, paraphrase, represent, translate, illustrate, give examples, classify, categorize, summarize, generalize, infer a logical conclusion (such as from examples given), predict, compare/contrast, match like ideas, explain, construct models	Compose & decompose numbers Evaluate an expression Locate points (grid/, number line) Represent math relationships in words pictures, or symbols Write simple sentences Select appropriate word for intended meaning Describe/explain how or why	Specify and explain relationships Give non-examples/examples Make and record observations Take notes; organize ideas/data Summarize results, concepts, ideas Make basic inferences or logical predictions from data or texts Identify main ideas or accurate generalizations	Explain, generalize, or connect ideas using supporting evidence Explain thinking when more than one response is possible Explain phenomena in terms of concepts Write full composition to meet specific purpose Identify themes	Explain how concepts or ideas specifically relate to other content domains or concepts Develop generalizations of the results obtained or strategies used and apply them to new problem situations
Apply Carry out or use a procedure in a given situation; carry out (apply to a familiar task), or use (apply) to an unfamiliar task	Follow simple/routine procedure (recipe-type directions) Solve a one-step problem Calculate, measure, apply a rule Apply an algorithm or formula (area, perimeter, etc.) Represent in words or diagrams a concept or relationship Apply rules or use resources to edit spelling, grammar, punctuation, conventions	Select a procedure according to task needed and perform it Solve routine problem applying multiple concepts or decision points Retrieve information from a table, graph, or figure and use it solve a problem requiring multiple steps Use models to represent concepts Write paragraph using appropriate organization; text structure, and signal words	Use concepts to solve non-routine problems Design investigation for a specific purpose or research question Conduct a designed investigation Apply concepts to solve non-routine problems Use reasoning, planning, and evidence Revise final draft for meaning or progression of ideas	Select or devise an approach among many alternatives to solve a novel problem Conduct a project that specifies a problem, identifies solution paths, solves the problem, and reports results Illustrate how multiple themes (historical, geographic, social) may be interrelated
Analyze Break into constituent parts, determine how parts relate, differentiate between relevant-irrelevant, distinguish, focus, select, organize, outline, find coherence, deconstruct (e.g., for bias or point of view)	Retrieve information from a table or graph to answer a question Identify or locate specific information contained in maps, charts, tables, graphs, or diagrams	Categorize, classify materials Compare/ contrast figures or data Select appropriate display data Organize or interpret (simple) data Extend a pattern Identify use of literary devices Identify text structure of paragraph Distinguish: relevant-irrelevant information; fact/opinion	Compare information within or across data sets or texts Analyze and draw conclusions from more complex data Generalize a pattern Organize/interpret data: complex graph Analyze author's craft, viewpoint, or potential bias	Analyze multiple sources of evidence or multiple works by the same author, or across genres, or time periods Analyze complex/abstract themes Gather, analyze, and organize information Analyze discourse styles
Evaluate Make judgments based on criteria, check, detect inconsistencies or fallacies, judge, critique			Cite evidence and develop a logical argument for concepts Describe, compare, and contrast solution methods Verify reasonableness of results Justify conclusions made	Gather, analyze, & evaluate relevancy & accuracy Draw & justify conclusions Apply understanding in a novel way, provide argument or justification for the application
Create Reorganize elements into new patterns/structures, generate, hypothesize, design, plan, construct, produce	Brainstorm ideas, concepts, or perspectives related to a topic or concept	Generate conjectures or hypotheses based on observations or prior knowledge	Synthesize information within one source or text Formulate an original problem, given a situation Develop a complex model for a given situation	Synthesize information across multiple sources or texts Design a model to inform and solve a real-world, complex, or abstract situation

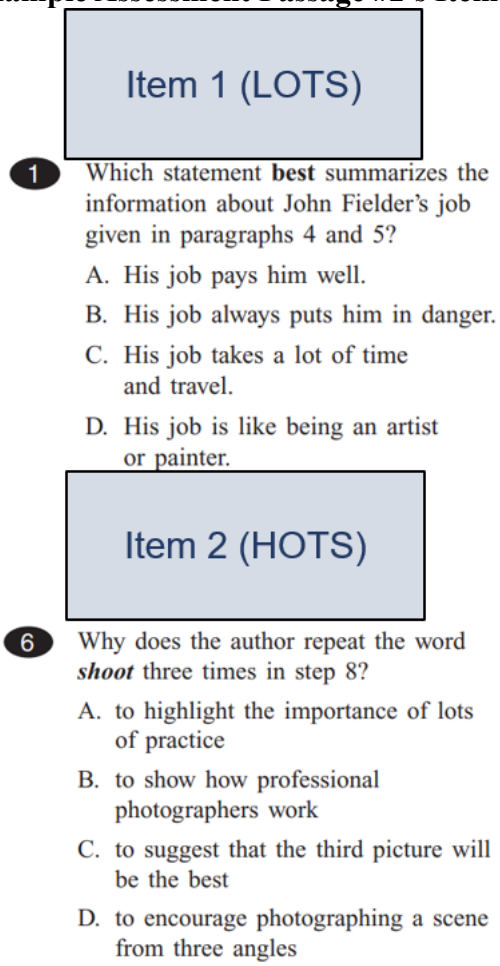
Note: The original Cognitive Rigor matrix is available from Dr. Hess's website, where permission to reproduce is given when authorship is fully cited.

Appendix Figure 2: Example Assessment Passage #1's Items (Along with Coding)



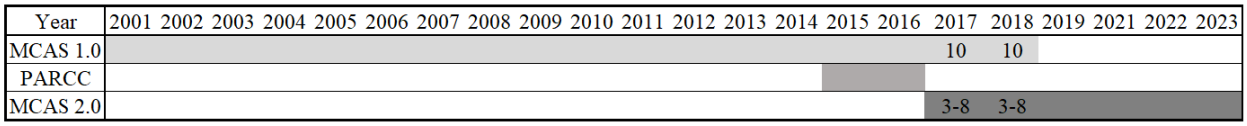
Note: These items are sourced from the Massachusetts Department of Education's publicly available 2004 Grade 7 assessment, which permits reproduction for non-commercial purposes. The items are based on a passage from *Island of the Blue Dolphins* by Scott O'Dell, which are not shown due to their copyright.

Appendix Figure 3: Example Assessment Passage #2's Items (Along with Coding)



Note: These items are sourced from the Massachusetts Department of Education's publicly available 2006 Grade 5 assessment, which permits reproduction for non-commercial purposes. The items are based on a passage from *Taking His Best Shots* by Claudia Cangilla McAdam, which are not shown due to their copyright.

Appendix Figure 4: Massachusetts Testing Regime Year Breakdown

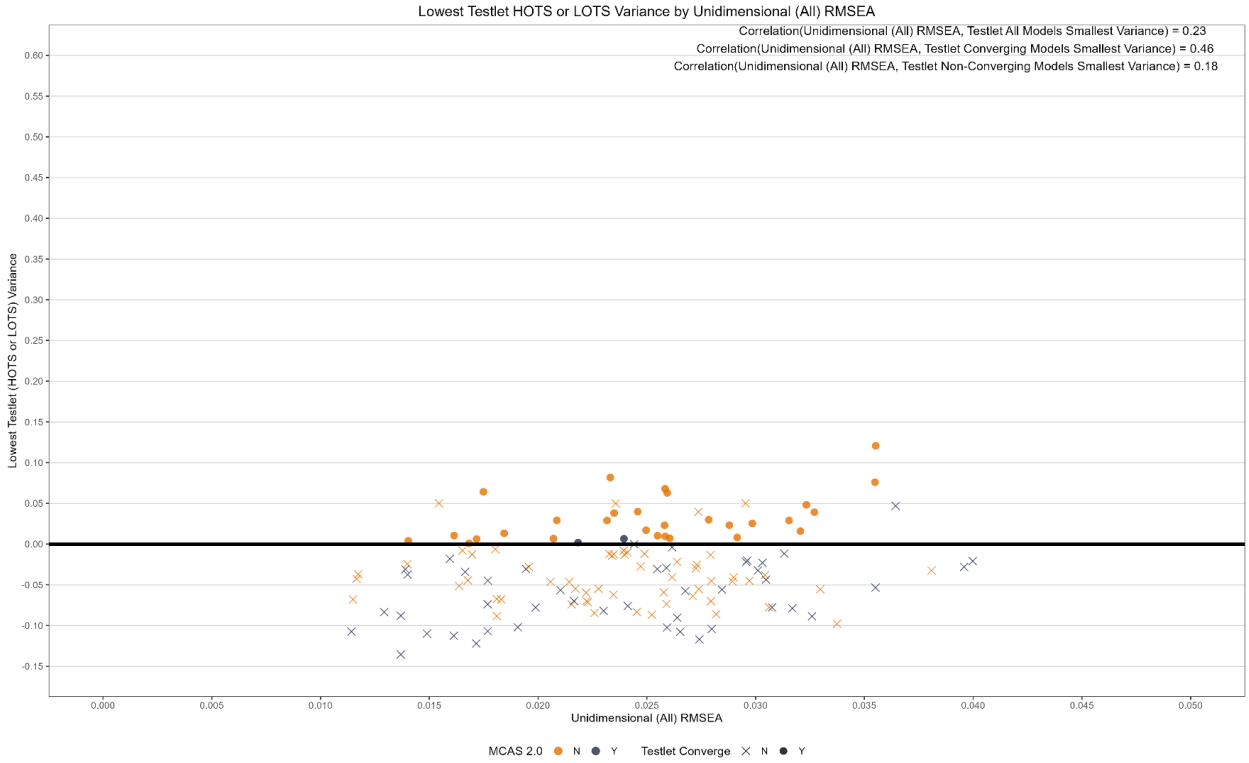


Massachusetts Comprehensive Assessment System (MCAS) has shifted from 2001 to 2023. While no assessment is identical, each era of assessment (e.g., MCAS 1.0) were built on similar standards and principles. MCAS 2.0 incorporated updated Common Core State Standards as well as shifted to using computers for exams. 3rd-8th grade assessments shifted to MCAS 2.0 starting in 2017, while 10th grade assessments shifted to MCAS 2.0 starting in 2019.

Appendix Figure 5: Massachusetts DOK Level Distribution in 2001 and 2023

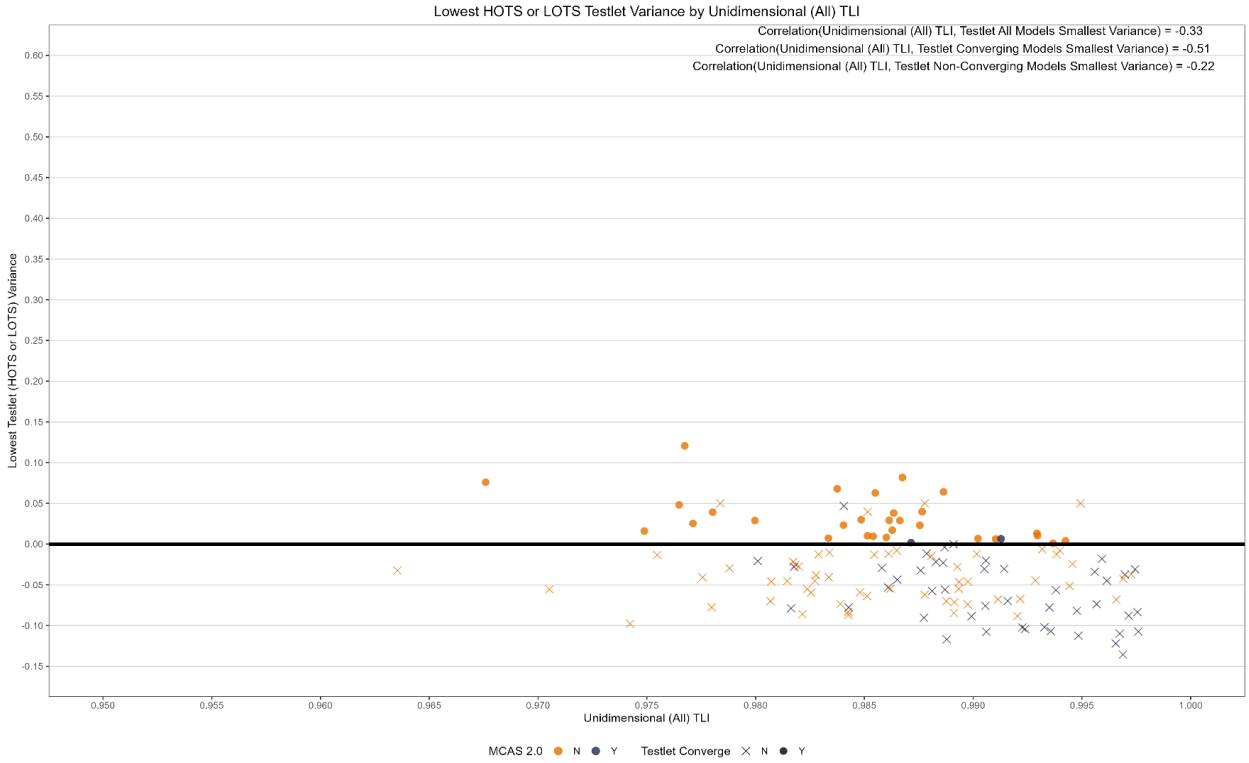
Note: These bar graphs represent the percentage of items that are classified into each DOK level across all assessment combinations in a year (2001 or 2023).

Appendix Figure 6: Lowest HOTS or LOTS Testlet Variance by Unidimensional (All) RMSEA Scatterplot



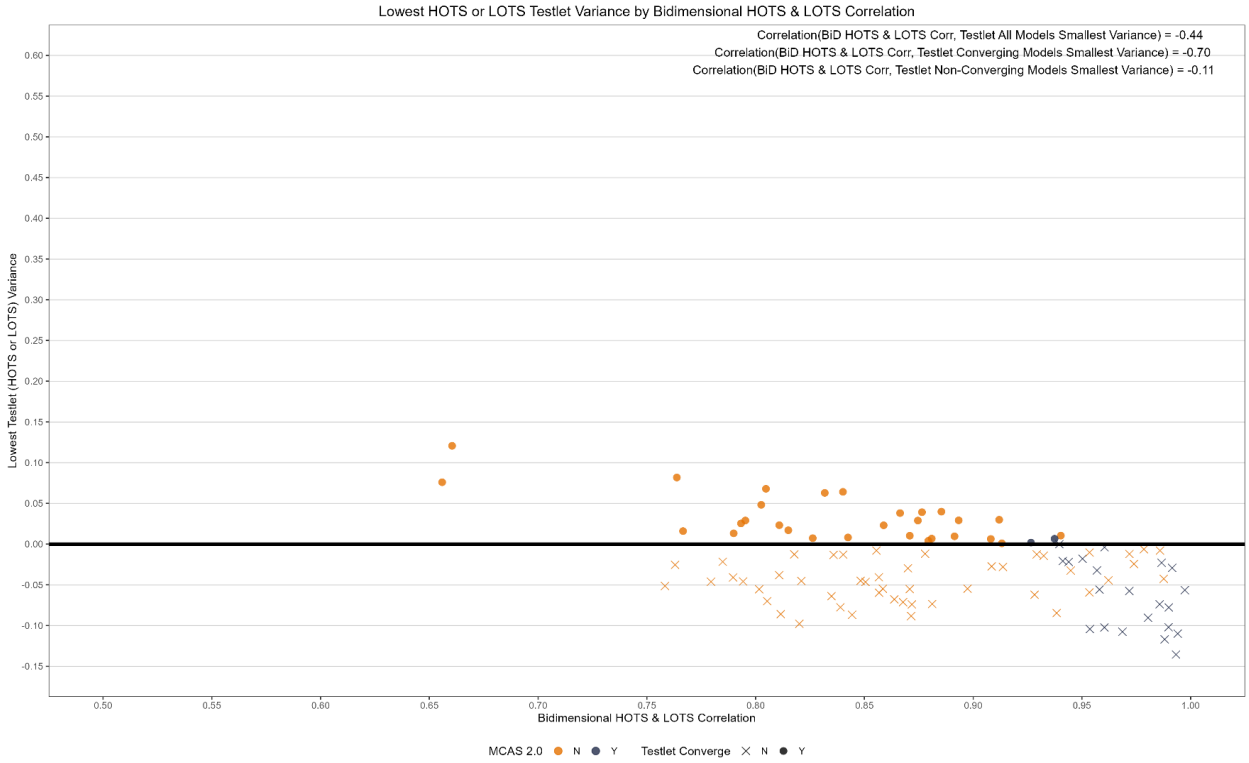
Note: This figure provides a visual representation of the testlet model's variance and the unidimensional model's Root Mean Squared Error of Approximation (RMSEA). Specifically, the x-axis displays the RMSEA of the unidimensional model, and the y-axis shows the variance of either the LOTS or HOTS factor (whichever is lowest) from the testlet model. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's HOTS or LOTS variance and the unidimensional RMSEA.

Appendix Figure 7: Lowest HOTS or LOTS Testlet Variance by Unidimensional (All) TLI Scatterplot



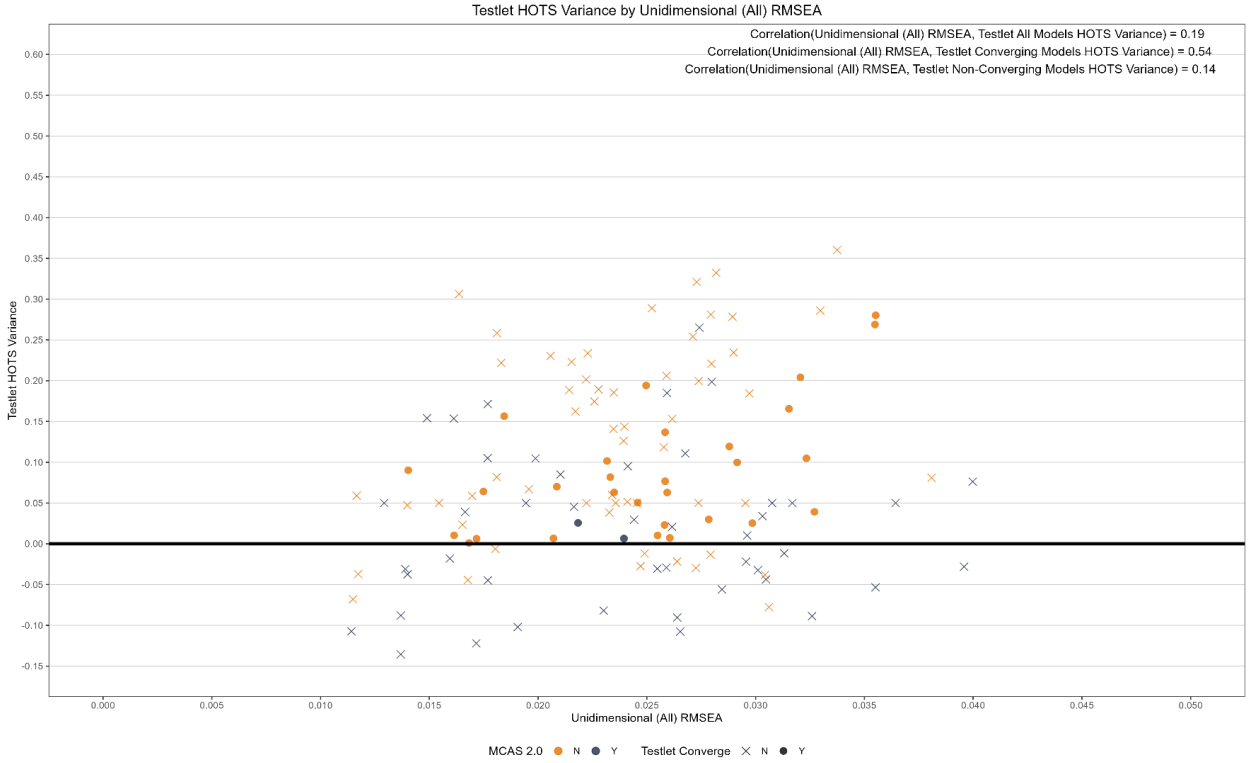
Note: This figure provides a visual representation of the testlet model's variance and the unidimensional model's Tucker Lewis Index (TLI). Specifically, the x-axis displays the TLI of the unidimensional model, and the y-axis shows the variance of either the LOTS or HOTS factor (whichever is lowest) from the testlet model. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's HOTS or LOTS variance and the unidimensional TLI.

Appendix Figure 8: Lowest HOTS or LOTS Testlet Variance by Bidimensional HOTS & LOTS Correlation Scatterplot



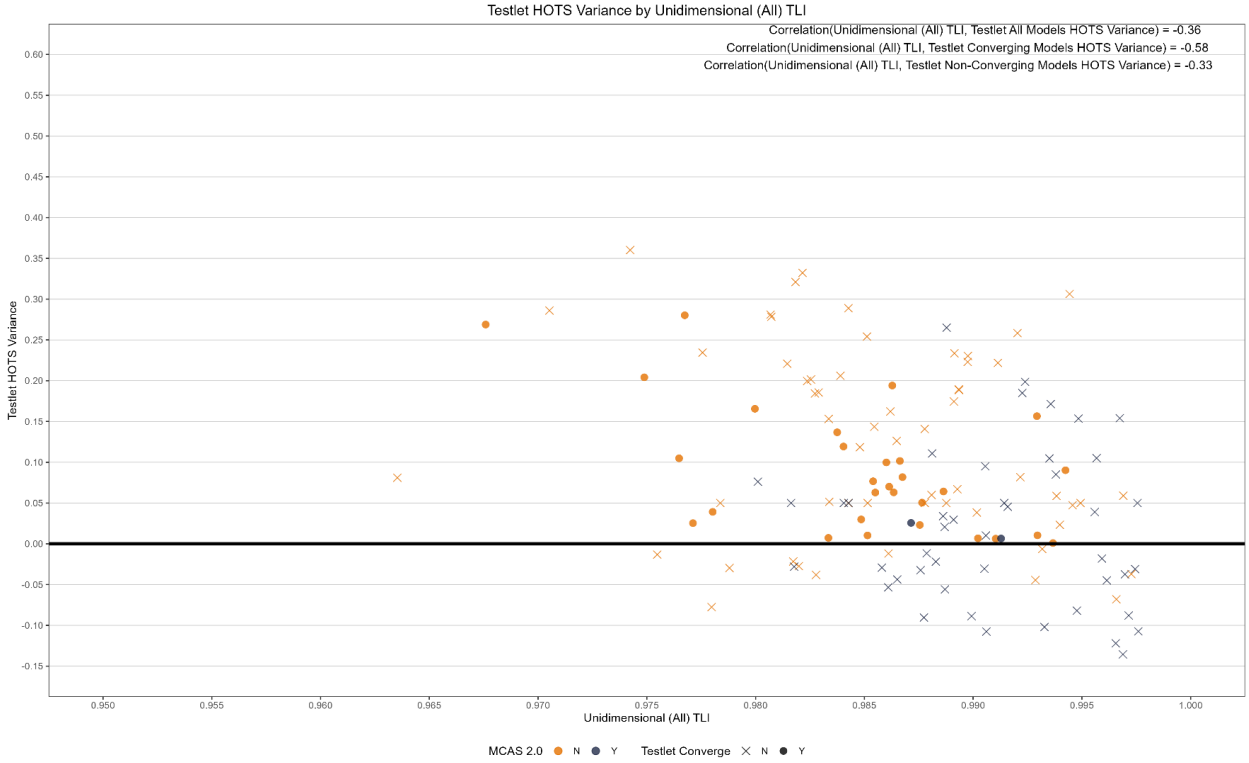
Note: This figure provides a visual representation of the testlet model's variance and the bidimensional model's HOTS and LOTS factor correlation. Specifically, the x-axis displays the HOTS and LOTS correlation of the bidimensional model, and the y-axis shows the variance of either the LOTS or HOTS factor (whichever is lowest) from the testlet model. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's HOTS or LOTS variance and the bidimensional HOTS and LOTS factor correlations.

Appendix Figure 9: HOTS Testlet Variance by Unidimensional (All) RMSEA Scatterplot



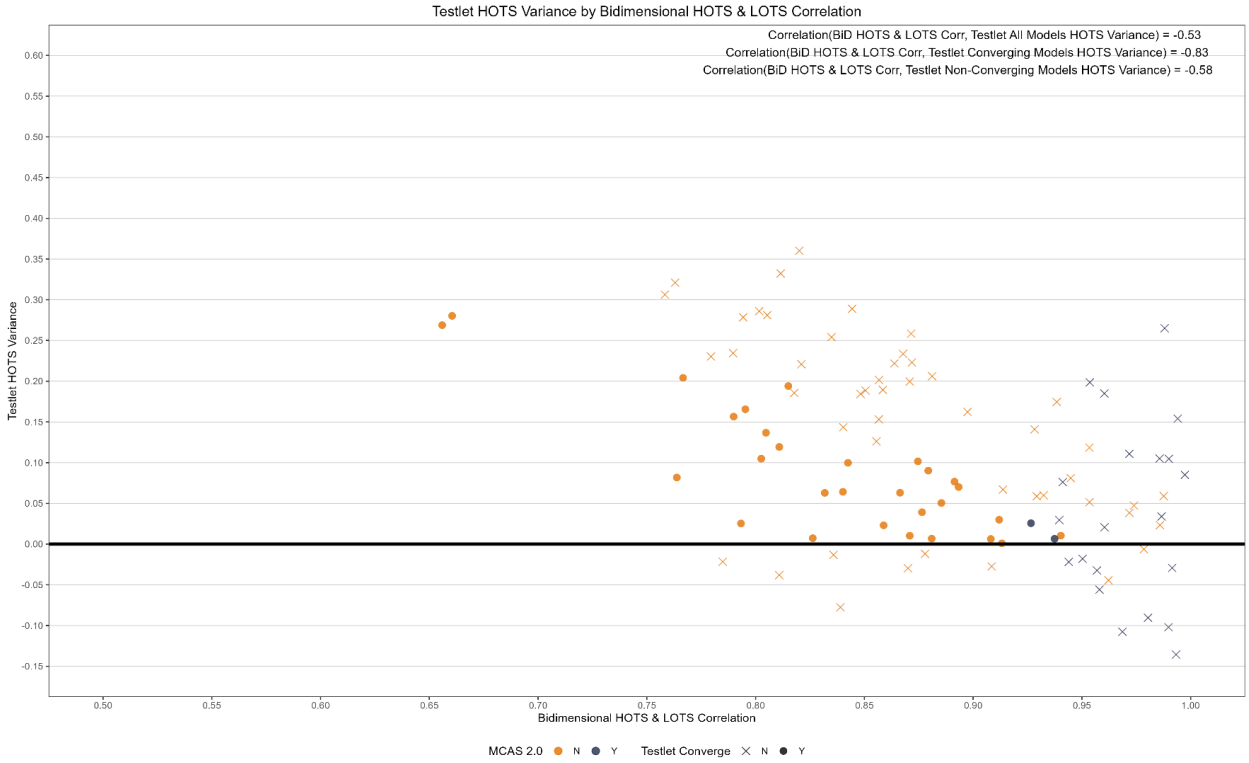
Note: This figure provides a visual representation of the testlet model's HOTS variance and the unidimensional model's Root Mean Squared Error of Approximation (RMSEA). Specifically, the x-axis displays the RMSEA of the unidimensional model, and the y-axis shows the HOTS variance from the testlet model. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's HOTS variance and the unidimensional RMSEA.

Appendix Figure 10: HOTS Testlet Variance by Unidimensional (All) TLI Scatterplot



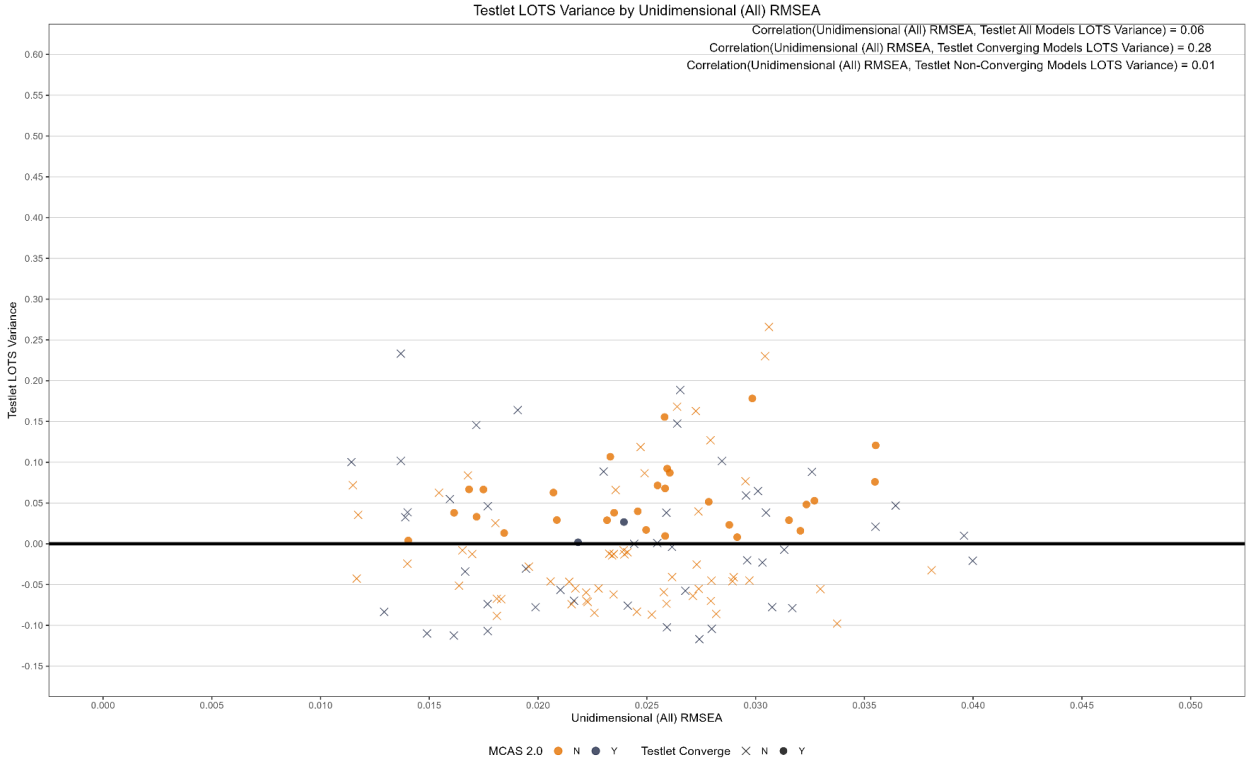
Note: This figure provides a visual representation of the testlet model's HOTS variance and the unidimensional model's Tucker Lewis Index (TLI). Specifically, the x-axis displays the TLI of the unidimensional model, and the y-axis shows the HOTS variance from the testlet model. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's HOTS variance and the unidimensional TLI.

Appendix Figure 11: HOTS Testlet Variance by Bidimensional HOTS & LOTS Correlation Scatterplot



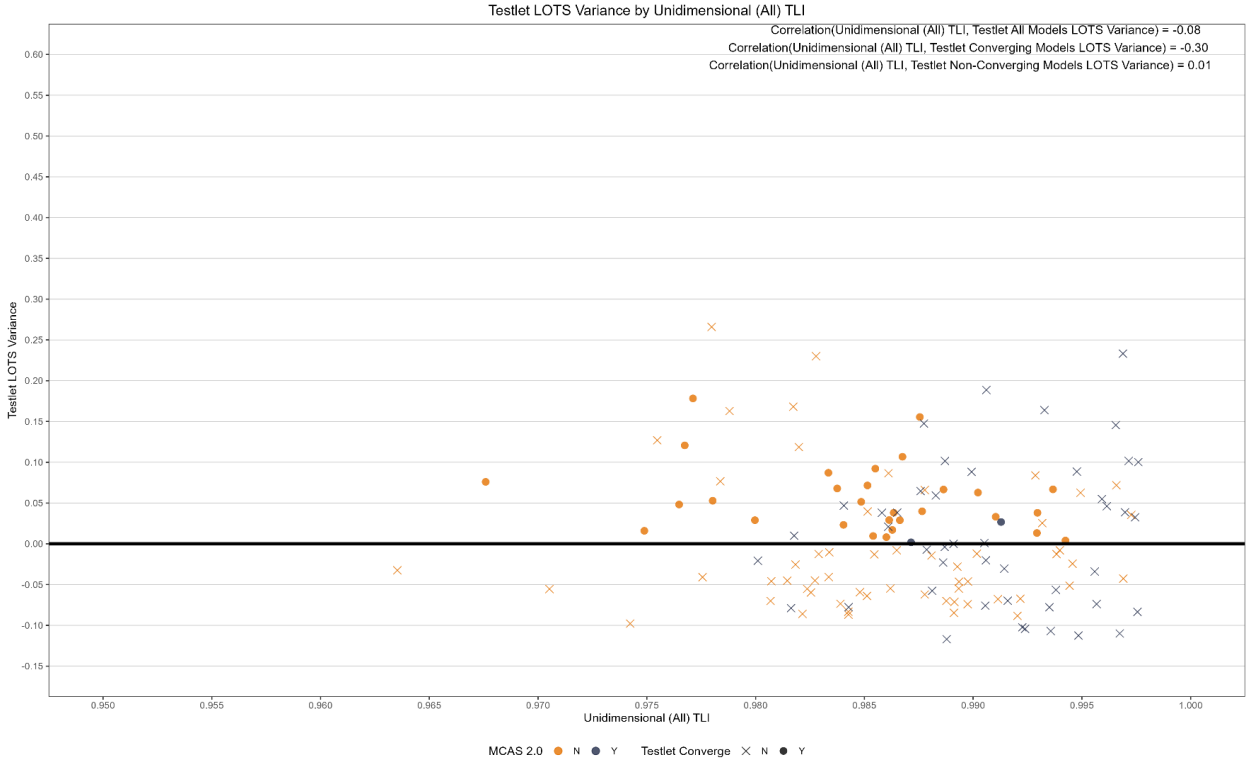
Note: This figure provides a visual representation of the testlet model's HOTS variance and the bidimensional model's HOTS and LOTS factor correlation. Specifically, the x-axis displays the HOTS and LOTS correlation of the bidimensional model, and the y-axis shows the HOTS variance from the testlet model. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's HOTS variance and the bidimensional HOTS and LOTS factor correlations.

Appendix Figure 12: LOTS Testlet Variance by Unidimensional (All) RMSEA Scatterplot



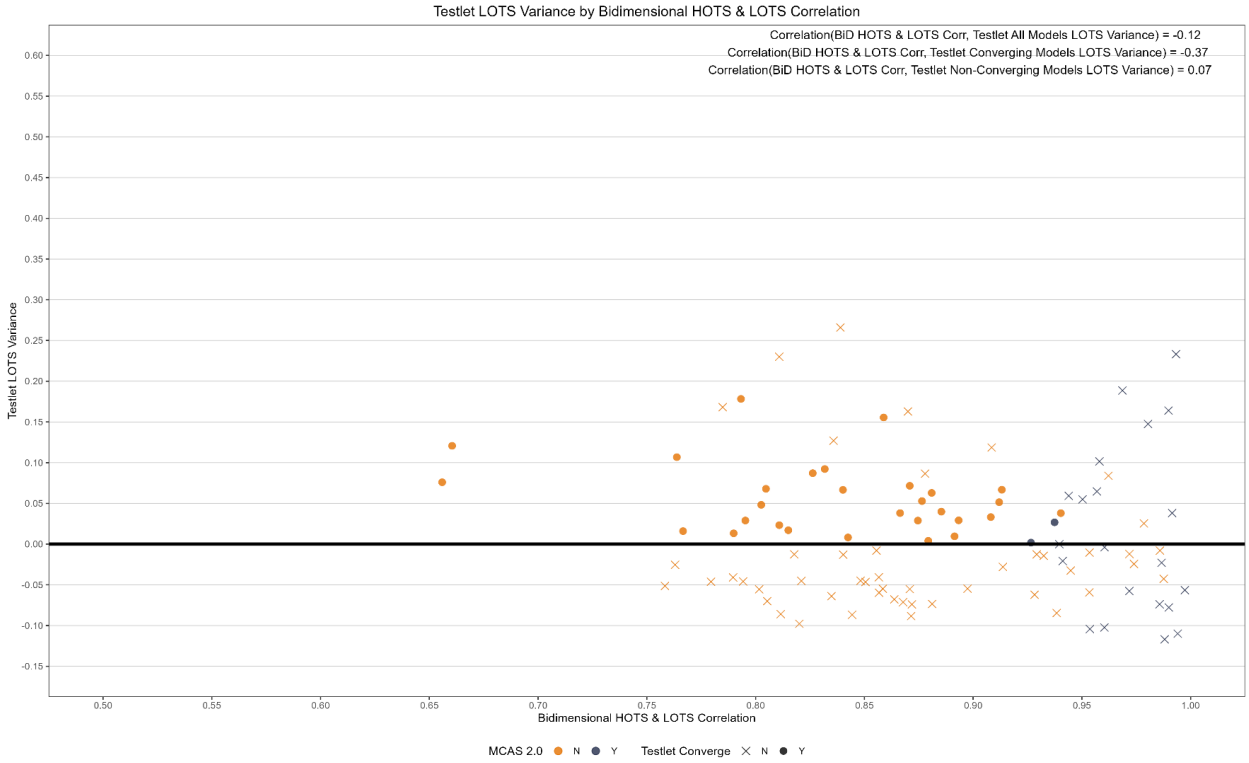
Note: This figure provides a visual representation of the testlet model's LOTS variance and the unidimensional model's Root Mean Squared Error of Approximation (RMSEA). Specifically, the x-axis displays the RMSEA of the unidimensional model, and the y-axis shows the LOTS variance from the testlet model. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's LOTS variance and the unidimensional RMSEA.

Appendix Figure 13: LOTS Testlet Variance by Unidimensional (All) TLI Scatterplot



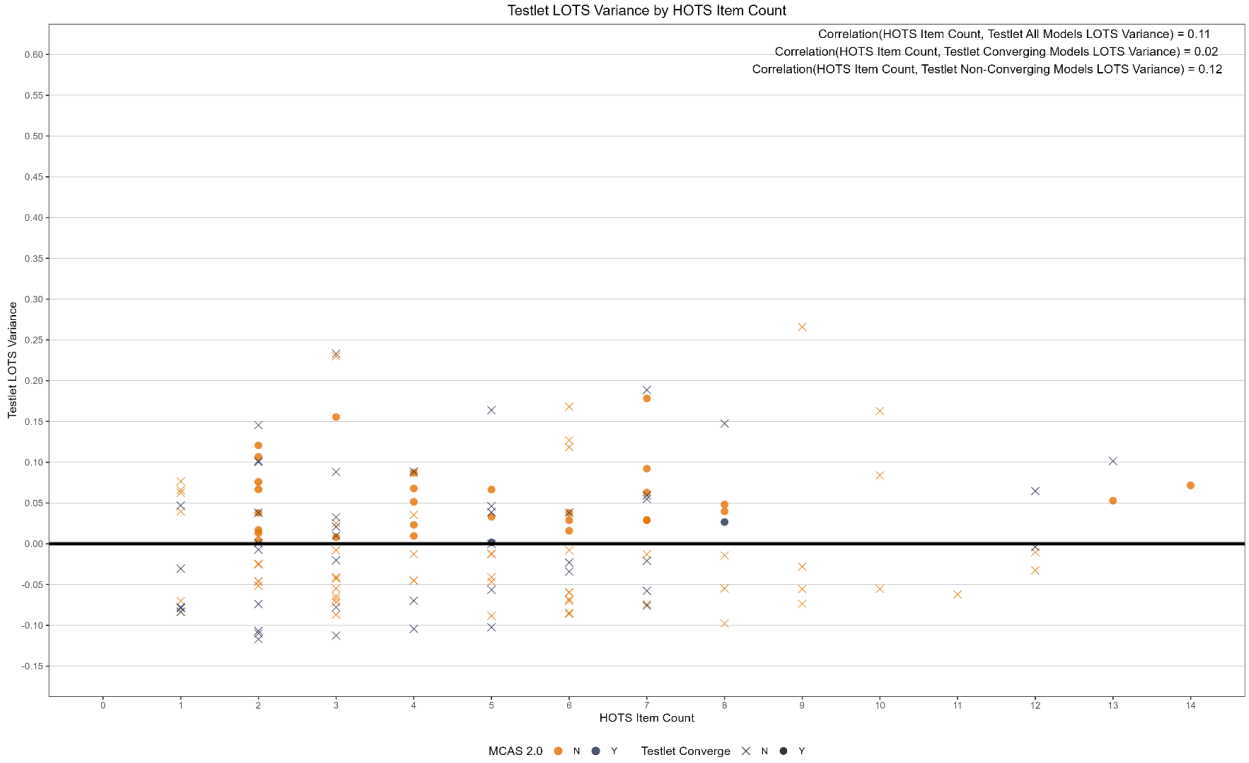
Note: This figure provides a visual representation of the testlet model's LOTS variance and the unidimensional model's Tucker Lewis Index (TLI). Specifically, the x-axis displays the TLI of the unidimensional model, and the y-axis shows the LOTS variance from the testlet model. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's LOTS variance and the unidimensional TLI.

Appendix Figure 14: LOTS Testlet Variance by Bidimensional HOTS & LOTS Correlation Scatterplot



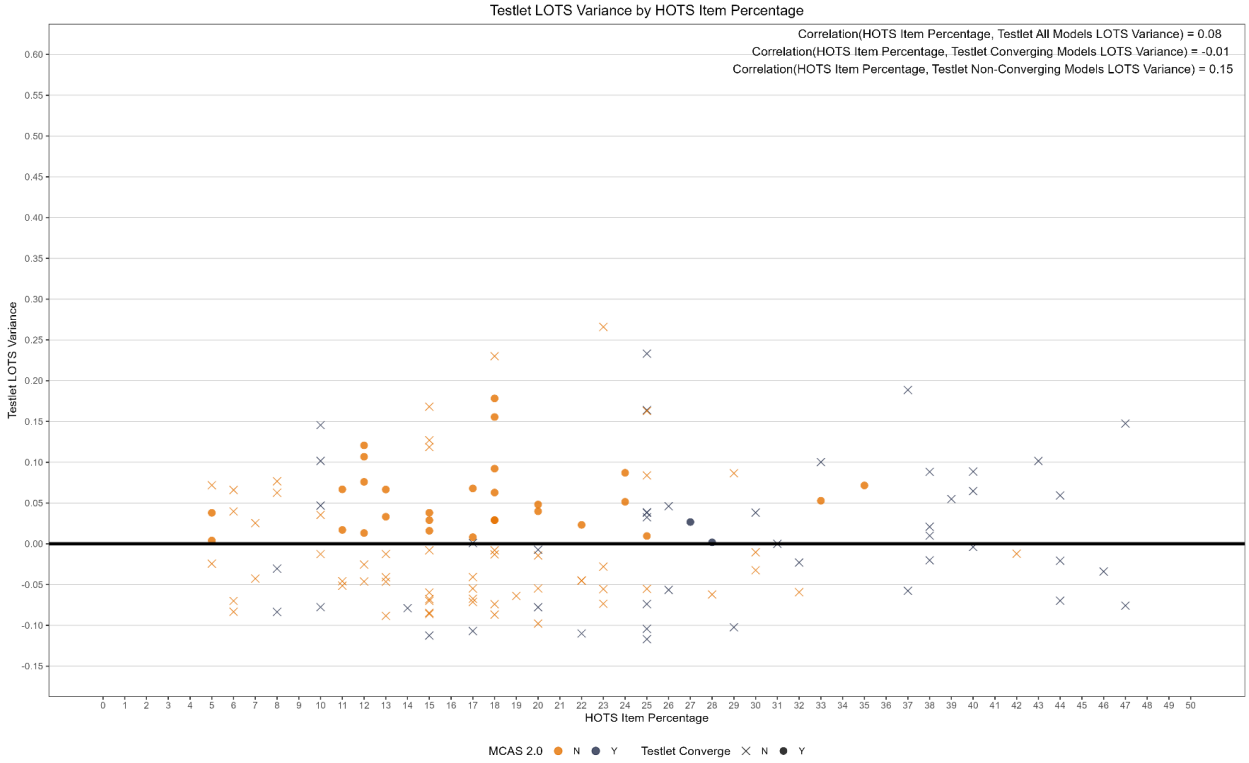
Note: This figure provides a visual representation of the testlet model's LOTS variance and the bidimensional model's HOTS and LOTS factor correlation. Specifically, the x-axis displays the HOTS and LOTS correlation of the bidimensional model, and the y-axis shows the LOTS variance from the testlet model. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's LOTS variance and the bidimensional HOTS and LOTS factor correlations.

Appendix Figure 15: LOTS Testlet Variance by HOTS Item Count Scatterplot



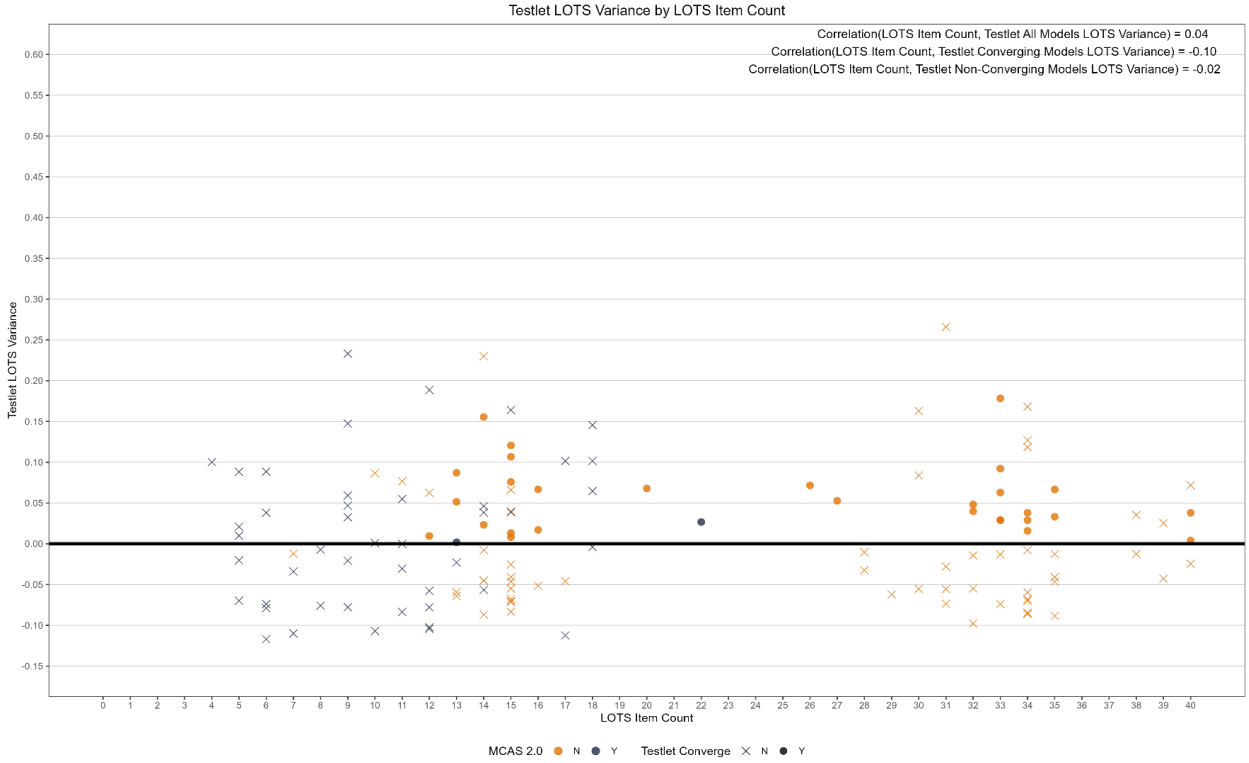
Note: This figure provides a visual representation of the testlet model's LOTS variance and the assessment's HOTS item count. Specifically, the x-axis displays the HOTS item count of the assessment, and the y-axis shows the LOTS variance from the testlet model. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's LOTS variance and the assessment's HOTS item count.

Appendix Figure 16: LOTS Testlet Variance by HOTS Item Percentage Scatterplot



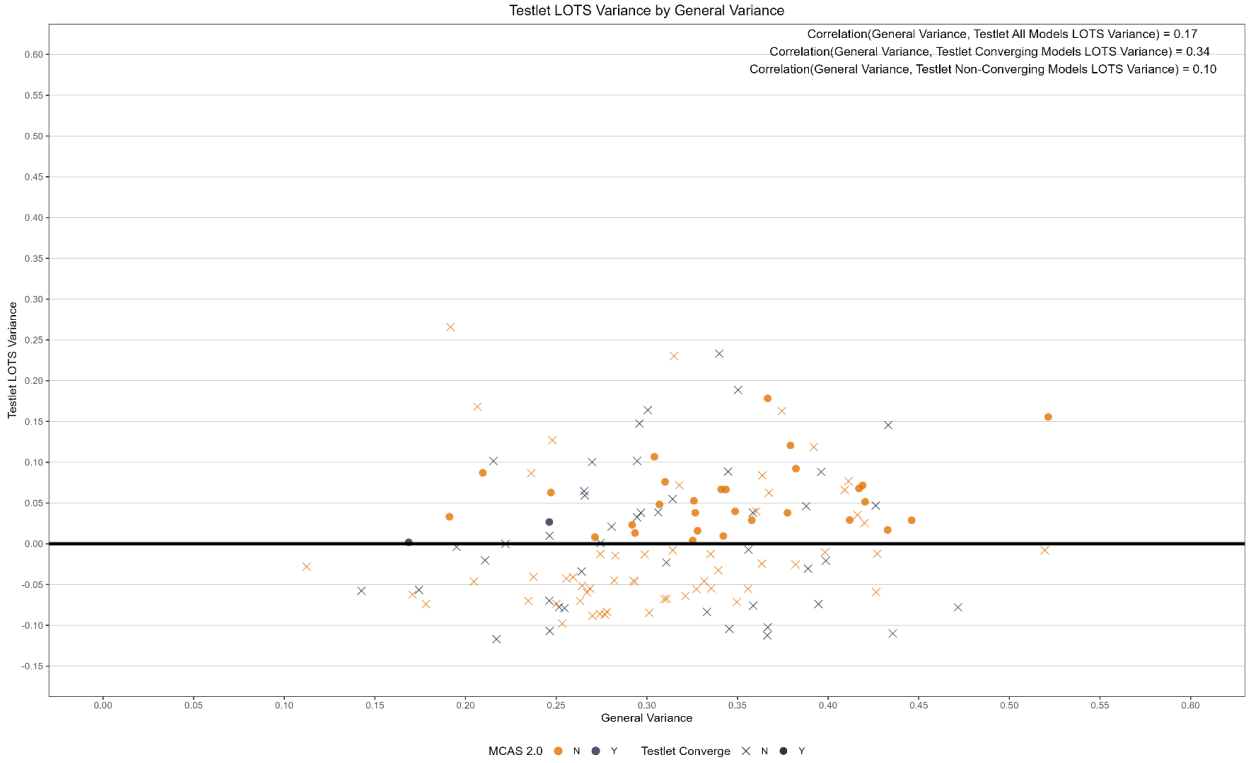
Note: This figure provides a visual representation of the testlet model's LOTS variance and the assessment's HOTS item percentage. Specifically, the x-axis displays the HOTS item percentage of the assessment, and the y-axis shows the LOTS variance from the testlet model. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's LOTS variance and the assessment's HOTS item percentage.

Appendix Figure 17: LOTS Testlet Variance by LOTS Item Count Scatterplot



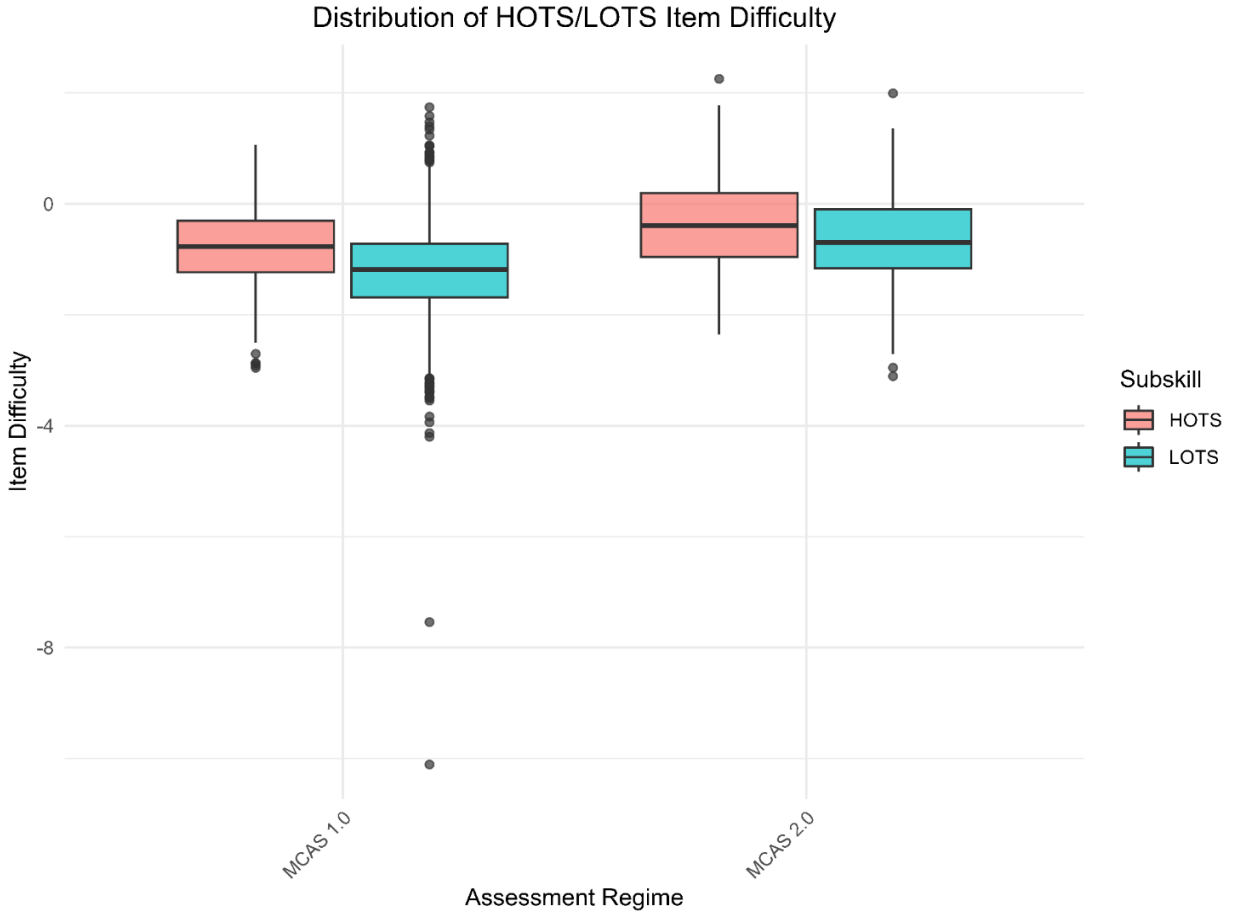
Note: This figure provides a visual representation of the testlet model's LOTS variance and the assessment's LOTS item count. Specifically, the x-axis displays the LOTS item count of the assessment, and the y-axis shows the LOTS variance from the testlet model. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's LOTS variance and the assessment's LOTS item count.

Appendix Figure 18: LOTS Testlet Variance by General ELA Variance Scatterplot



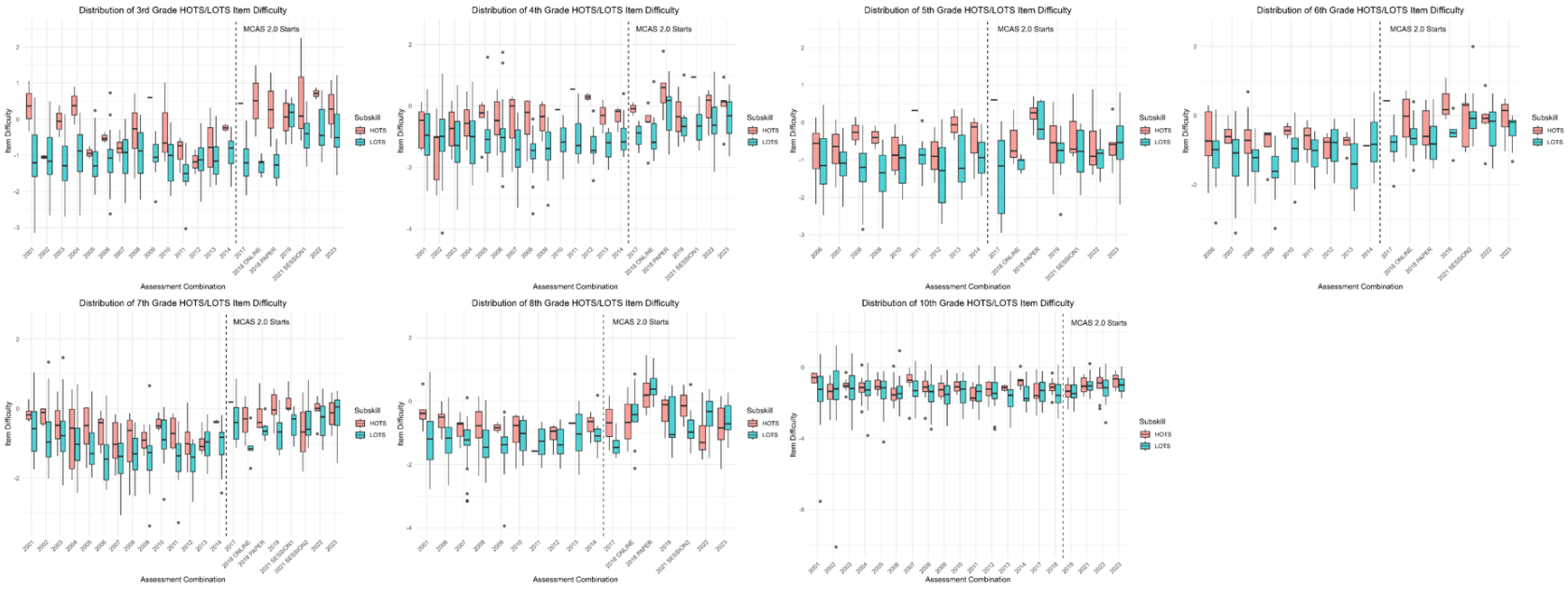
Note: This figure provides a visual representation of the testlet model's LOTS and General ELA variance. Specifically, the x-axis displays the variance of the General ELA factor, and the y-axis shows the LOTS variance. The graph also features indicators for the test regime (MCAS 1.0 or MCAS 2.0) along with whether the testlet model converged. The top right describes the association between the testlet model's LOTS variance and the testlet model's General ELA variance.

Appendix Figure 19: Item Difficulty



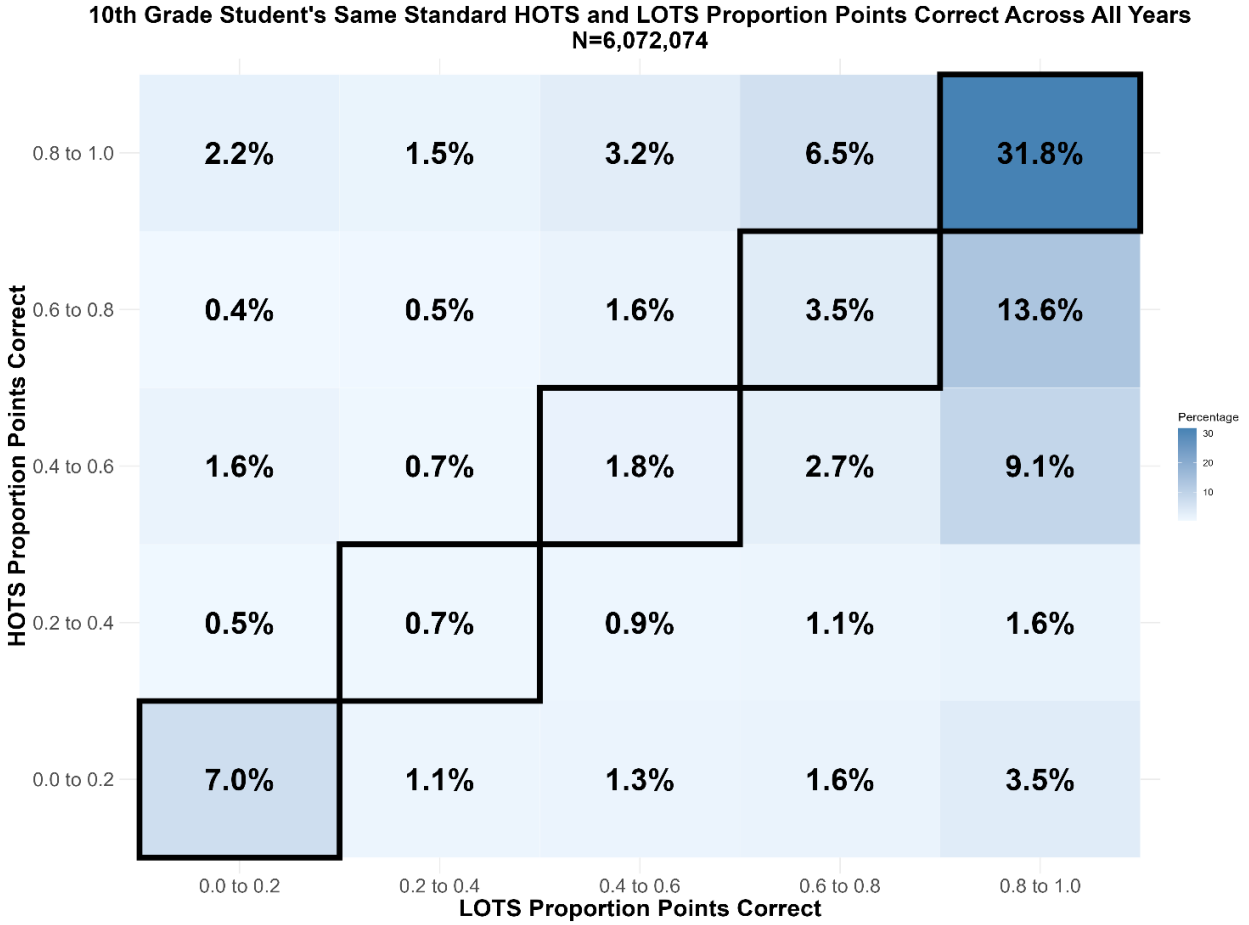
Note: Item difficulties are generated using Item Response Theory Three-Parameter Logistic and Graded Response Models with all items (including those without text released). Then the items with depths of knowledge 1 and 2 are grouped as lower-order thinking skills (LOTS) and depths of knowledge 3 and 4 are grouped as higher-order thinking skills (HOTS) for MCAS 1.0 and MCAS 2.0. The item difficulties are pooled across all grades.

Appendix Figure 20: Item Difficulty by Grade



Note: Item difficulties are generated using Item Response Theory Three-Parameter Logistic and Graded Response Models with all items (including those without text released). Then the items with depths of knowledge 1 and 2 are grouped as lower-order thinking skills (LOTS) and depths of knowledge 3 and 4 are grouped as higher-order thinking skills (HOTS) for MCAS 1.0 and MCAS 2.0. The item difficulties are separated for each grade.

Appendix Figure 21: 10th Grade Student's Same Standard HOTS and LOTS Proportion Points Correct Across All Years



Note: We compute 10th students' LOTS and HOTS performance (proportion of points earned) for each ELA standard (e.g., Student 1 may have 1.00/1.00 points proportion earned for LOTS and 0.50/1.00 points proportion earned for HOTS for Standard 1). The y-axis represents the HOTS point proportion earned for each standard, and the x-axis represents the LOTS point proportion received. Each bin shows the percentage of 10th grade students that earned each combination of LOTS and HOTS points. For example in the top right bin, ~31.8% of student-standard observations received all LOTS points and also all HOTS points; however in the top left bin, only ~2.2% of students received almost no LOTS points but nearly all HOTS points.

Appendix Table 1: State ELA Assessment DOK Breakdown From Blueprints (When Available)

	Assessment Blueprints (Contemporary) - ELA HOTS Percentage									
	Grades									
State	3	4	5	6	7	8	9	10	11	12
AZ	31%-44%	31%-44%	31%-44%	31%-44%	31%-44%	31%-44%	31%-44%	31%-44%	31%-44%	31%-44%
AR	16%-31%	16%-31%	16%-31%	16%-31%	16%-31%	16%-31%	16%-31%	16%-31%	16%-31%	16%-31%
CA			22			44				
CT			22			44				
DE			22			44				
GA	25%-45%	35%-55%	35%-55%	35%-55%	35%-55%	35%-55%				
HI			22			44				
ID			22			44				
IA	30%-45%	30%-45%	30%-45%	30%-45%	30%-45%	30%-45%	30%-45%	30%-45%	30%-45%	30%-45%
MI			22			44				
MN	>10%	>10%	>10%	>10%	>10%	>10%		>10%		
MO			22			44				
MT			22			44				
NV			22			44				
NM	20%-40%	20%-40%	20%-40%	20%-40%	20%-40%	20%-40%				
NC	0	5%-10%	10%-25%	18%-40%	18%-40%	18%-40%				
ND	18%-29%	18%-29%	18%-29%	18%-29%	18%-29%	18%-29%	18%-29%	18%-29%		
OH	5%-13%	5%-13%	5%-13%	5%-13%	5%-13%	5%-13%	5%-13%	5%-13%	5%-13%	5%-13%
OR			22			44				
SC	8%-35%	8%-35%	8%-35%	8%-35%	8%-35%	8%-35%	8%-35%	8%-35%	8%-35%	8%-35%
SD			22			44				
UT	24%-34%	24%-34%	24%-34%	24%-34%	24%-34%	24%-34%				
VT	20%-40%	20%-40%	20%-40%	20%-40%	20%-40%	20%-40%	20%-40%			
WA			22			44				
WV	6%-12%	6%-12%	6%-12%	6%-15%	6%-15%	6%-15%				
WI	38.5	15.4	23.1	23.1	28.2	33.3				
WY	28%-41%	28%-41%	28%-41%	28%-41%	28%-41%	28%-41%	28%-41%	28%-41%		

Note: These depth of knowledge (DOK) breakdowns were obtained from the most recent publicly available blueprints of ELA assessments from each state. The DOK 3 or 4, or HOTS, percentage is then reported. The Smarter Balanced Assessment Consortium does not release its current DOK breakdown, so we report the estimated HOTS breakdown based on Doorey and Polikoff (2016).

Appendix Table 2: Individual Assessment Combination Fit Statistics for Each Measurement Model

	Unidimensional (All Items)				Unidimensional (Released Items)				Bidimensional (Released Items)				Testlet (Released Items)			
Assessment	RMSEA	SRMSR	TLI	CFI	RMSEA	SRMSR	TLI	CFI	RMSEA	SRMSR	TLI	CFI	RMSEA	SRMSR	TLI	CFI
Year: 2001, Grade: 3	0.01403	0.02413	0.99424	0.99452	0.01403	0.02413	0.99424	0.99452	0.01372	0.02395	0.9945	0.99477	0.01372	0.02395	0.9945	0.99477
Year: 2002, Grade: 3	0.01614	0.02516	0.99296	0.9933	0.01614	0.02516	0.99296	0.9933	0.01607	0.02512	0.99302	0.99337	0.01607	0.02512	0.99302	0.99337
Year: 2003, Grade: 3	0.01149	0.02227	0.99658	0.99675	0.01149	0.02227	0.99658	0.99675								
Year: 2004, Grade: 3	0.01399	0.02396	0.99457	0.99483	0.01399	0.02396	0.99457	0.99483	0.01399	0.02396	0.99457	0.99484				
Year: 2005, Grade: 3	0.01166	0.02096	0.9969	0.99705	0.01166	0.02096	0.9969	0.99705	0.01166	0.02096	0.9969	0.99706				
Year: 2006, Grade: 3	0.01804	0.02824	0.99318	0.99351	0.01804	0.02824	0.99318	0.99351	0.01803	0.02823	0.99318	0.99352				
Year: 2007, Grade: 3	0.01173	0.02103	0.99727	0.9974	0.01173	0.02103	0.99727	0.9974								
Year: 2008, Grade: 3	0.01697	0.02574	0.99383	0.99413	0.01697	0.02574	0.99383	0.99413	0.01675	0.02593	0.99399	0.99429				
Year: 2009, Grade: 3	0.01544	0.02517	0.99494	0.99518	0.01321	0.02029	0.99659	0.99716								
Year: 2010, Grade: 3	0.0181	0.02872	0.99217	0.99256	0.02169	0.03322	0.99074	0.99183								
Year: 2011, Grade: 3	0.01652	0.02688	0.99398	0.99428	0.01105	0.01852	0.99741	0.99773	0.01103	0.01858	0.99742	0.99776				
Year: 2012, Grade: 3	0.01844	0.02706	0.99293	0.99329	0.02349	0.02961	0.99015	0.99138	0.01782	0.02555	0.99433	0.99508	0.01782	0.02555	0.99433	0.99508
Year: 2013, Grade: 3	0.01637	0.02674	0.99443	0.99471	0.02	0.02823	0.99208	0.99301	0.01707	0.02594	0.99423	0.99494				
Year: 2014, Grade: 3	0.01683	0.02818	0.99367	0.99399	0.01489	0.02252	0.99574	0.99624	0.01409	0.02211	0.99618	0.99666	0.01409	0.02211	0.99618	0.99666
Year: 2017, Grade: 3	0.01944	0.02604	0.99143	0.99214	0.01327	0.02073	0.99698	0.99753								
Year: 2018, Grade: 3 PAPER	0.01489	0.02147	0.99674	0.99699	0.0085	0.01542	0.999	0.99925	0.00872	0.01536	0.99895	0.99924				
Year: 2018, Grade: 3 ONLINE	0.01768	0.02122	0.99568	0.99605	0.00948	0.01346	0.9987	0.99907	0.00956	0.01334	0.99868	0.9991				
Year: 2019, Grade: 3	0.0259	0.02946	0.98581	0.98704	0.01407	0.01582	0.9948	0.99628	0.01445	0.01581	0.99451	0.99627				
Year: 2021, Grade: 3 SESSION1	0.0139	0.01791	0.99745	0.99791	0.0139	0.01791	0.99745	0.99791								
Year: 2022, Grade: 3	0.01369	0.01916	0.99715	0.99734	0.01376	0.01955	0.997	0.99732								
Year: 2023, Grade: 3	0.01716	0.02153	0.99656	0.99678	0.01729	0.02266	0.99684	0.99717								
Year: 2001, Grade: 4	0.01717	0.0284	0.99104	0.9915	0.01717	0.0284	0.99104	0.9915	0.01611	0.02807	0.99211	0.99253	0.01611	0.02807	0.99211	0.99253
Year: 2002, Grade: 4	0.01749	0.02728	0.98864	0.98922	0.01749	0.02728	0.98864	0.98922	0.01446	0.02578	0.99223	0.99264	0.01446	0.02578	0.99223	0.99264
Year: 2003, Grade: 4	0.02471	0.03715	0.98199	0.98291	0.02471	0.03715	0.98199	0.98291	0.02389	0.03693	0.98317	0.98405				
Year: 2004, Grade: 4	0.01677	0.02617	0.99286	0.99323	0.01677	0.02617	0.99286	0.99323	0.01645	0.02624	0.99314	0.9935				
Year: 2005, Grade: 4	0.0181	0.02555	0.99204	0.99244	0.0181	0.02555	0.99204	0.99244	0.01487	0.02401	0.99463	0.99491				
Year: 2006, Grade: 4	0.02086	0.02948	0.98614	0.98685	0.02086	0.02948	0.98614	0.98685	0.01929	0.02934	0.98815	0.98877	0.01929	0.02934	0.98815	0.98877
Year: 2007, Grade: 4	0.02397	0.03242	0.98546	0.9862	0.02397	0.03242	0.98546	0.9862	0.01959	0.03163	0.99029	0.9908				
Year: 2008, Grade: 4	0.02639	0.03288	0.98173	0.98267	0.02639	0.03288	0.98173	0.98267	0.01844	0.02808	0.99108	0.99155				
Year: 2009, Grade: 4	0.02605	0.03185	0.98335	0.9842	0.03101	0.03533	0.97698	0.97986	0.02552	0.03411	0.98441	0.98648	0.02552	0.03411	0.98441	0.98648
Year: 2010, Grade: 4	0.02357	0.02992	0.98778	0.9884	0.0138	0.02195	0.99588	0.99643								
Year: 2011, Grade: 4	0.02454	0.03059	0.98426	0.98507	0.01378	0.0224	0.99497	0.99564								
Year: 2012, Grade: 4	0.02057	0.02832	0.98976	0.99028	0.0225	0.02745	0.98751	0.98907	0.01464	0.0224	0.99471	0.99541				
Year: 2013, Grade: 4	0.02278	0.02807	0.98934	0.98989	0.02509	0.02906	0.98727	0.98877	0.02108	0.02761	0.99101	0.99213				
Year: 2014, Grade: 4	0.02581	0.03117	0.98755	0.98819	0.03254	0.03339	0.98393	0.98594	0.02681	0.03139	0.98909	0.99054	0.02681	0.03139	0.98909	0.99054
Year: 2017, Grade: 4	0.02548	0.02666	0.99052	0.99131	0.01404	0.01839	0.99745	0.99791								
Year: 2018, Grade: 4 PAPER	0.02184	0.04987	0.98715	0.98813	0.01337	0.04682	0.99453	0.99518	0.01251	0.0468	0.99521	0.99581	0.01251	0.0468	0.99521	0.99581
Year: 2018, Grade: 4 ONLINE	0.02442	0.02751	0.9891	0.99005	0.02037	0.02405	0.98977	0.99114	0.01945	0.02417	0.99068	0.992				
Year: 2019, Grade: 4	0.02741	0.02718	0.98878	0.98976	0.01468	0.01704	0.99691	0.99779	0.015	0.01702	0.99677	0.99781				
Year: 2021, Grade: 4 SESSION1	0.01292	0.01789	0.99755	0.998	0.01292	0.01789	0.99755	0.998								
Year: 2022, Grade: 4	0.01612	0.02149	0.99484	0.99518	0.01344	0.01872	0.99712	0.99743	0.01347	0.0187	0.99711	0.99743				
Year: 2023, Grade: 4	0.01906	0.02662	0.99328	0.99371	0.02058	0.03043	0.99348	0.99417	0.02062	0.03042	0.99346	0.99418				
Year: 2006, Grade: 5	0.02347	0.02901	0.98777	0.9884	0.02347	0.02901	0.98777	0.9884	0.02231	0.02935	0.98895	0.98953				
Year: 2007, Grade: 5	0.02154	0.02701	0.98975	0.99027	0.02154	0.02701	0.98975	0.99027	0.01741	0.02523	0.99331	0.99366				
Year: 2008, Grade: 5	0.02349	0.03016	0.98635	0.98705	0.02349	0.03016	0.98635	0.98705	0.02065	0.0287	0.98945	0.99001	0.02065	0.0287	0.98945	0.99001
Year: 2009, Grade: 5	0.02332	0.02991	0.98675	0.98743	0.02322	0.03038	0.98586	0.98763	0.01384	0.02419	0.99498	0.99565	0.01384	0.02419	0.99498	0.99565
Year: 2010, Grade: 4	0.02357	0.02992	0.98778	0.9884	0.0138	0.02195	0.99588	0.99643								
Year: 2011, Grade: 4	0.02454	0.03059	0.98426	0.98507	0.01378	0.0224	0.99497	0.99564								
Year: 2012, Grade: 4	0.02057	0.02832	0.98976	0.99028	0.0225	0.02745	0.98751	0.98907	0.01464	0.0224	0.99471	0.99541				
Year: 2013, Grade: 4	0.02278	0.02807	0.98934	0.98989	0.02509	0.02906	0.98727	0.98877	0.02108	0.02761	0.99101	0.99213				
Year: 2014, Grade: 4	0.02581	0.03117	0.98755	0.98819	0.03254	0.03339	0.98393	0.98594	0.02681	0.03139	0.98909	0.99054	0.02681	0.03139	0.98909	0.99054
Year: 2017, Grade: 4	0.02548	0.02666	0.99052	0.99131	0.01404	0.01839	0.99745	0.99791								
Year: 2018, Grade: 4 PAPER	0.02184	0.04987	0.98715	0.98813	0.01337	0.04682	0.99453	0.99518	0.01251	0.0468	0.99521	0.99581	0.01251	0.0468	0.99521	0.99581
Year: 2018, Grade: 4 ONLINE	0.02442	0.02751	0.9891	0.99005	0.02037	0.02405	0.98977	0.99114	0.01945	0.02417	0.99068	0.992				
Year: 2019, Grade: 4	0.02741	0.02718	0.98878	0.98976	0.01468	0.01704	0.99691	0.99779	0.015	0.01702	0.99677	0.99781				
Year: 2021, Grade: 4 SESSION1	0.01292	0.01789	0.99755	0.998	0.01292	0.01789	0.99755	0.998								
Year: 2022, Grade: 4	0.01612	0.02149	0.99484	0.99518	0.01344	0.01872	0.99712	0.99743	0.01347	0.0187	0.99711	0.99743				
Year: 2023, Grade: 4	0.01906	0.02662	0.99328	0.99371	0.02058	0.03043	0.99348	0.99417	0.02062	0.03042	0.99346	0.99418				
Year: 2006, Grade: 5	0.02347	0.02901	0.98777	0.9884	0.02347	0.02901	0.98777	0.9884	0.02231	0.02935	0.98895	0.98953				
Year: 2007, Grade: 5	0.02154	0.02701	0.98975	0.99027	0.02154	0.02701	0.98975	0.99027	0.01741	0.02523	0.99331	0.99366				
Year: 2008, Grade: 5	0.02349	0.03016	0.98635	0.98705	0.02349	0.03016	0.98635	0.98705	0.02065	0.0287	0.98945	0.99001	0.02065	0.0287	0.98945	0.99001
Year: 2009, Grade: 5	0.02332	0.02991	0.98675	0.98743	0.02322	0.03038	0.98586	0.98763	0.01384	0.02419	0.99498	0.99565	0.01384	0.02419	0.99498	0.99565
Year: 2010, Grade: 5	0.02584	0.03118	0.9854	0.98615	0.02675	0.02973	0.98401	0.98614	0.02448	0.02968	0.98661	0.98851	0.02448	0.02968	0.98661	0.98851
Year: 2011, Grade: 5	0.02223	0.02804	0.98877	0.98934	0.01752	0.02593	0.99379	0.99457	0.01759	0.02593	0.99373	0.99456				
Year: 2012, Grade: 5	0.02616	0.03223	0.98336	0.98421	0.0286	0.03451	0.98358	0.98551	0.02407	0.03477	0.98837	0.98981				
Year: 2013, Grade: 5	0.02497	0.03233	0.98628	0.98699	0.02613	0.032	0.98616	0.98779	0.01767	0.02618	0.99367	0.99446	0.01767	0.02618	0.99367	0.99446
Year: 2014, Grade: 5	0.02523	0.03056	0.98427	0.98507	0.02849	0.03141	0.98186	0.98413	0.02277	0.03003	0.98841	0.98995				
Year: 2017, Grade: 5	0.03169	0.0352	0.98163	0.98316	0.02072	0.02682	0.98718	0.99145	0.02153	0.02682	0.98616	0.99143				
Year: 2018, Grade: 5 PAPER	0.02165	0.0467	0.99159	0.99224	0.01695	0.03927	0.99579	0.99684								
Year: 2018, Grade: 5 ONLINE	0.02961	0.03	0.99059	0.99141	0.01603	0.01696	0.99728	0.99806								

Year: 2019, Grade: 5	0.0264	0.02966	0.98774	0.98876	0.02337	0.02615	0.98914	0.9905	0.02335	0.02639	0.98916	0.99059				
Year: 2021, Grade: 5 SESSION1	0.01369	0.01893	0.99689	0.99745	0.01369	0.01893	0.99689	0.99745	0.01382	0.01895	0.99683	0.99745				
Year: 2022, Grade: 5	0.01769	0.02425	0.99615	0.9964	0.0118	0.01683	0.99846	0.99863								
Year: 2023, Grade: 5	0.02103	0.02409	0.9938	0.99422	0.01608	0.02274	0.99603	0.99647	0.01614	0.02274	0.996	0.99647				
Year: 2006, Grade: 6	0.0183	0.02471	0.99114	0.99159	0.0183	0.02471	0.99114	0.99159	0.01442	0.02308	0.9945	0.99479				
Year: 2007, Grade: 6	0.02393	0.03354	0.98649	0.98718	0.02393	0.03354	0.98649	0.98718	0.02015	0.03244	0.99042	0.99093				
Year: 2008, Grade: 6	0.02725	0.03323	0.9788	0.97989	0.02725	0.03323	0.9788	0.97989	0.02466	0.03361	0.98264	0.98355				
Year: 2009, Grade: 6	0.02797	0.03476	0.98145	0.98241	0.02813	0.03482	0.98076	0.98302	0.02148	0.03213	0.98878	0.99018				
Year: 2010, Grade: 6	0.02729	0.03427	0.98183	0.98276	0.03149	0.03641	0.98362	0.98567	0.01918	0.02936	0.99392	0.99473				
Year: 2011, Grade: 6	0.03043	0.03661	0.98278	0.98366	0.03237	0.03549	0.98486	0.98676	0.02358	0.03159	0.99196	0.99303				
Year: 2012, Grade: 6	0.02228	0.02875	0.98914	0.9897	0.0232	0.0272	0.98989	0.99108	0.01873	0.02616	0.99342	0.99423				
Year: 2013, Grade: 6	0.02584	0.03222	0.98375	0.98458	0.03172	0.03936	0.97747	0.97943	0.02267	0.03302	0.98849	0.98953	0.02267	0.03302	0.98849	0.98953
Year: 2014, Grade: 6	0.02737	0.03242	0.98515	0.98591	0.01699	0.02573	0.99422	0.99499								
Year: 2017, Grade: 6	0.03644	0.03539	0.98406	0.98538	0.01356	0.01833	0.9968	0.99751								
Year: 2018, Grade: 6 PAPER	0.02593	0.02972	0.99226	0.99288	0.02164	0.02667	0.99486	0.99551	0.02076	0.02635	0.99528	0.9959				
Year: 2018, Grade: 6 ONLINE	0.02798	0.02867	0.99239	0.99305	0.0216	0.0228	0.99558	0.99617	0.02007	0.02235	0.99618	0.99672				
Year: 2019, Grade: 6	0.03551	0.03653	0.9861	0.98731	0.01658	0.01738	0.99529	0.99663								
Year: 2021, Grade: 6 SESSION2	0.014	0.01985	0.99698	0.9973	0.014	0.01985	0.99698	0.9973								
Year: 2022, Grade: 6	0.03032	0.02941	0.98862	0.98938	0.01996	0.02693	0.99469	0.99528	0.01995	0.0269	0.99469	0.99531				
Year: 2023, Grade: 6	0.02412	0.02443	0.99056	0.99119	0.01229	0.01627	0.99615	0.9967								
Year: 2001, Grade: 7	0.02899	0.03686	0.97756	0.97871	0.02899	0.03686	0.97756	0.97871	0.02337	0.03352	0.98542	0.98618				
Year: 2002, Grade: 7	0.02348	0.03047	0.9829	0.98377	0.02348	0.03047	0.9829	0.98377	0.01804	0.0269	0.9899	0.99043				
Year: 2003, Grade: 7	0.0234	0.03126	0.98809	0.9887	0.0234	0.03126	0.98809	0.9887	0.02246	0.03114	0.98902	0.9896				
Year: 2004, Grade: 7	0.02317	0.03253	0.98663	0.98732	0.02317	0.03253	0.98663	0.98732	0.02029	0.03189	0.98975	0.99029	0.02029	0.03189	0.98975	0.99029
Year: 2005, Grade: 7	0.02259	0.03082	0.98912	0.98968	0.02259	0.03082	0.98912	0.98968	0.022	0.03047	0.98968	0.99022				
Year: 2006, Grade: 7	0.03206	0.04049	0.97488	0.97617	0.03206	0.04049	0.97488	0.97617	0.02156	0.0337	0.98864	0.98924	0.02156	0.0337	0.98864	0.98924
Year: 2007, Grade: 7	0.02458	0.03393	0.98765	0.98829	0.02458	0.03393	0.98765	0.98829	0.02131	0.03342	0.99072	0.99121	0.02131	0.03342	0.99072	0.99121
Year: 2008, Grade: 7	0.02594	0.03365	0.9855	0.98625	0.02594	0.03365	0.9855	0.98625	0.01891	0.03153	0.99229	0.9927	0.01891	0.03153	0.99229	0.9927
Year: 2009, Grade: 7	0.03549	0.04525	0.96759	0.96925	0.03876	0.04696	0.95137	0.95745	0.01527	0.02496	0.99246	0.99346	0.01527	0.02496	0.99246	0.99346
Year: 2010, Grade: 7	0.02711	0.0347	0.98512	0.98588	0.03516	0.04072	0.97982	0.98251	0.02862	0.03805	0.98663	0.98853				
Year: 2011, Grade: 7	0.02971	0.03553	0.98273	0.98361	0.02782	0.0344	0.97835	0.9809	0.02366	0.03341	0.98433	0.98628				
Year: 2012, Grade: 7	0.02578	0.03332	0.9848	0.98558	0.02162	0.02747	0.99072	0.99175	0.02094	0.02754	0.99129	0.99231				
Year: 2013, Grade: 7	0.0249	0.03	0.98611	0.98682	0.02718	0.02986	0.98448	0.98687	0.02309	0.02936	0.9888	0.99064				
Year: 2014, Grade: 7	0.02915	0.03523	0.98601	0.98673	0.02916	0.03214	0.9871	0.98861	0.02307	0.02976	0.99192	0.99293	0.02307	0.02976	0.99192	0.99293
Year: 2017, Grade: 7	0.03076	0.03298	0.98428	0.98559	0.01479	0.01824	0.99405	0.99537								
Year: 2018, Grade: 7 PAPER	0.02301	0.03407	0.99477	0.99517	0.01503	0.02824	0.99796	0.99842								
Year: 2018, Grade: 7 ONLINE	0.03259	0.03238	0.98993	0.9908	0.01272	0.01392	0.99839	0.99885								
Year: 2019, Grade: 7	0.03998	0.04195	0.9801	0.98183	0.03564	0.03325	0.9859	0.98778	0.03464	0.03287	0.98668	0.98856				
Year: 2021, Grade: 7 SESSION1	0.01665	0.02189	0.99559	0.99632	0.01665	0.02189	0.99559	0.99632								
Year: 2021, Grade: 7 SESSION2	0.01142	0.0171	0.99759	0.99786	0.01192	0.01542	0.9948	0.99654								
Year: 2022, Grade: 7	0.03048	0.03134	0.98651	0.98738	0.02084	0.02698	0.99255	0.99333								
Year: 2023, Grade: 7	0.01988	0.02175	0.99351	0.99394	0.01272	0.01629	0.9975	0.99786	0.01274	0.01627	0.99749	0.99787				
Year: 2001, Grade: 8	0.02143	0.02846	0.98935	0.98993	0.02143	0.02846	0.98935	0.98993	0.01581	0.02417	0.9942	0.99452				
Year: 2006, Grade: 8	0.02071	0.02745	0.99022	0.99072	0.02071	0.02745	0.99022	0.99072	0.0176	0.02647	0.99293	0.99331	0.0176	0.02647	0.99293	0.99331
Year: 2007, Grade: 8	0.03154	0.03708	0.97997	0.981	0.03154	0.03708	0.97997	0.981	0.0227	0.03322	0.98962	0.99017	0.0227	0.03322	0.98962	0.99017
Year: 2008, Grade: 8	0.03061	0.03942	0.97798	0.97911	0.03061	0.03942	0.97798	0.97911	0.02555	0.03732	0.98465	0.98546				
Year: 2009, Grade: 8	0.03552	0.04153	0.97674	0.97793	0.04236	0.04496	0.9647	0.96911	0.01714	0.02741	0.99422	0.99499	0.01714	0.02741	0.99422	0.99499
Year: 2010, Grade: 8	0.02784	0.03422	0.98485	0.98563	0.02597	0.03232	0.98558	0.98739	0.02425	0.03291	0.98743	0.98909	0.02425	0.03291	0.98743	0.98909
Year: 2011, Grade: 8	0.02954	0.03686	0.97838	0.97949	0.01485	0.02096	0.99467	0.99564								
Year: 2012, Grade: 8	0.02879	0.0355	0.98404	0.98486	0.03591	0.03868	0.97774	0.98035	0.02639	0.03536	0.98798	0.98947	0.02639	0.03536	0.98798	0.98947
Year: 2013, Grade: 8	0.02894	0.03739	0.98072	0.98171	0.02953	0.0352	0.98385	0.98565	0.01925	0.02929	0.99313	0.99394				
Year: 2014, Grade: 8	0.02327	0.02923	0.99017	0.99067	0.03126	0.03497	0.98918	0.99115	0.0312	0.03544	0.98922	0.99135				
Year: 2017, Grade: 8	0.03132	0.03377	0.98786	0.98888	0.01274	0.015	0.9985	0.99884								
Year: 2018, Grade: 8 PAPER	0.01595	0.05028	0.99592	0.99624	0.01499	0.04862	0.9959	0.99638	0.01415	0.04839	0.99635	0.9968				
Year: 2018, Grade: 8 ONLINE	0.02956	0.03295	0.98828	0.9893	0.02462	0.0275	0.99022	0.99152	0.02365	0.02752	0.99098	0.99225				
Year: 2019, Grade: 8	0.03958	0.04059	0.98177	0.98336	0.01482	0.0171	0.99535	0.99668								
Year: 2021, Grade: 8 SESSION2	0.01768	0.02506	0.99357	0.99424	0.00896	0.01349	0.99858	0.99884								
Year: 2022, Grade: 8	0.02653	0.02669	0.99061	0.99123	0.01518	0.02112	0.99602	0.99646	0.01469	0.02093	0.99627	0.99671				
Year: 2023, Grade: 8	0.02677	0.02973	0.98812	0.98891	0.01795	0.02555	0.9932	0.99395	0.01771	0.02528	0.99338	0.99415				
Year: 2001, Grade: 10	0.02793	0.03776	0.97548	0.97677	0.02793	0.03776	0.97548	0.97677	0.02319	0.03232	0.98309	0.98401				
Year: 2002, Grade: 10	0.02221	0.03688	0.98254	0.98344	0.02221	0.03688	0.98254	0.98344	0.01858	0.03579	0.98778	0.98842				
Year: 2003, Grade: 10	0.02172	0.02998	0.9862	0.9869	0.02172	0.02998	0.9862	0.9869	0.01962	0.02962	0.98874	0.98933				
Year: 2004, Grade: 10	0.01957	0.02785	0.98927	0.98982	0.01957	0.02785	0.98927	0.98982	0.01786	0.02759	0.99107	0.99154				
Year: 2005, Grade: 10	0.02411	0.03241	0.98339	0.98424	0.02411	0.03241	0.98339	0.98424	0.02371	0.03265	0.98393	0.98477				
Year: 2006, Grade: 10	0.02549	0.03455	0.98514	0.9859	0.02549	0.03455	0.98514	0.9859	0.02361	0.03484	0.98726	0.98793	0.02361	0.03484	0.98726	0.98793
Year: 2007, Grade: 10	0.0259	0.03421	0.9839	0.98473	0.0259	0.03421	0.9839	0.98473	0.02301	0.03379	0.9873	0.98797				
Year: 2008, Grade: 10	0.03809	0.05059	0.96353	0.9654	0.03809	0.05059	0.96353	0.9654	0.0377	0.05025	0.96426	0.96614				
Year: 2009, Grade: 10	0.03298	0.04446	0.97052	0.97203	0.03298	0.04446	0.97052	0.97203	0.02568	0.03979	0.98212	0.98306				
Year: 2010, Grade: 10	0.02739	0.03485	0.98237	0.98327	0.02739	0.03485	0.98237	0.98327	0.02341	0.03342	0.98712	0.9878				
Year: 2011, Grade: 10	0.03233	0.0416	0.97648	0.97769	0.03233	0.0416	0.97648	0.97769	0.02373	0.03792	0.98734	0.988	0.02373	0.03792	0.98734	0.988
Year: 2012, Grade: 10	0.02819	0.03689	0.98215	0.98307	0.02819	0.03689	0.98215	0.98307	0.01967	0.03299	0.99131	0.99176				
Year: 2013, Grade: 10	0.02794	0.03826	0.98069	0.98168	0.02794	0.03826	0.98069	0.98168	0.01856	0.03099	0.99148	0.99192				
Year: 2014, Grade: 10	0.02985	0.03797	0.97712	0.97829	0.02985	0.03797	0.97712	0.97829	0.02099	0.03101	0.98868	0.98928	0.02099	0.03101	0.98868	0.98928
Year: 2017, Grade: 10	0.0327	0.04177	0.97803	0.97915	0.0327	0.04177	0.97803	0.97915	0.02924	0.04105	0.98243	0.98336	0.02924	0.04105	0.98243	0.98336
Year: 2018, Grade: 10	0.03374	0.04309	0.97423	0.97556	0.03374	0.04309	0.97423	0.97556	0.0267	0.03841	0.98387	0.98471				
Year: 2019, Grade: 10	0.02395	0.02578	0.99128	0.99189	0.02395	0.02578	0.99128	0.99189	0.02251	0.02606	0.9923	0.99285	0.02251	0.02606	0.9923	0.99285
Year: 2021, Grade: 10	0.02845	0.02898	0.9887	0.98948	0.02845	0.02898	0.9887	0.98948	0.02789	0.02939	0.98915	0.98992				
Year: 2022, Grade: 10	0.02616	0.0284	0.98869	0.98947	0.02616	0.0284	0.98869	0.98947	0.02571	0.02855	0.98908	0.98985				
Year: 2023, Grade: 10	0.03011	0.02955	0.98758	0.98844	0.03011	0.02955	0.98758	0.98844	0.02953	0.02972	0.98805	0.9889				

Note: The table documents the Root Mean Squared Error of Approximation (RMSEA), Standardized Root Mean Squared Residuals (SRMR), Comparative Fit Index (CFI), and the

Tucker-Lewis Index (TLI) for each individual assessment combination (grade-year-modality-session) and measurement model. All blank spaces indicate that the measurement model for that particular assessment combination does not align with the internal structure of the data and therefore the fit statistics are meaningless and not shown.

Appendix Table 3: Bidimensional Measurement Model Comparison Results Using RMSEA Confidence Intervals

Bidimensional (Released Items)	**Assessments** (N)	**Converged Models** (N)	**Met Hu & Bentler (1999) Good Fit** (% of Converged Models)	**Fit > Unidimensional (All Items)** (% of Converged Models)	**Fit > Unidimensional (Released Items)** (% of Converged Models)	**Fit > Unidimensional (Both All & Released)** (% of Converged Models)
MCAS 1.0	86	77	100%	74%	77%	69%
3rd	14	10	100%	20%	20%	0%
4th	14	12	100%	75%	92%	75%
5th	9	8	100%	75%	75%	75%
6th	9	8	100%	75%	88%	75%
7th	14	14	100%	93%	93%	86%
8th	10	9	100%	89%	78%	78%
10th	16	16	100%	81%	81%	81%

Note: This table replicates Table 7 but uses the lower 90% confidence interval for the unidimensional model RMSEA and the upper 90% confidence for the bidimensional model RMSEA instead. Since Massachusetts's public data provides student-level item performance for all items, we are able to estimate a unidimensional measurement model for using all items. In addition, we estimate a version with only released assessment items to create a similar comparison to the bidimensional and testlet models which use only released assessment items. The columns of the table then report out the number of assessment combinations (grade-year-modality/session), the number of models that converge without errors (e.g., negative variances or impossible correlations), and the percentage of models that meet Hu's and Bentler's (1999) 'good fit' criteria. The subsequents columns of the table compare the fit statistics between the bidimensional models that converge to both unidimensional counterparts. The values represent the percentage of models that provide a superior fit across all fit statistics.

Appendix Table 4: Bidimensional Measurement Model Comparison Results Using Chi-Squared Difference Tests

Bidimensional (Released Items)	**Assessments** (N)	**Converged Models** (N)	**Fit > Unidimensional (Released Items)** (% of Converged Models)
MCAS 1.0	86	77	100
3rd	14	10	100
4th	14	12	100
5th	9	8	100
6th	9	8	100
7th	14	14	100
8th	10	9	100
10th	16	16	100

Note: This table replicates Table 7 but uses the chi-squared difference test instead. Since Massachusetts's public data provides student-level item performance for all items, we are able to estimate a unidimensional measurement model for using all items. In addition, we estimate a version with only released assessment items to create a similar comparison to the bidimensional and testlet models which use only released assessment items. The columns of the table then report out the number of assessment combinations (grade-year-modality/session), the number of models that converge without errors (e.g., negative variances or impossible correlations), and the percentage of models that meet Hu's and Bentler's (1999) 'good fit' criteria. The subsequents columns of the table compare the fit statistics between the bidimensional models that converge to both unidimensional counterparts. The values represent the percentage of models that provide a superior fit across all fit statistics.

Appendix Table 5: Continuous Measurement Model Comparison Results

Bidimensional (Released Items)	**Assessments** (N)	**Converged Models** (N)	**Fit > Unidimensional (All Items)** (% of Converged Models)	**Fit > Unidimensional (Released Items)** (% of Converged Models)	**Fit > Unidimensional (Both All & Released)** (% of Converged Models)
MCAS 1.0	86	86	100	98	98
3rd	14	14	100	100	100
4th	14	14	100	93	93
5th	9	9	100	89	89
6th	9	9	100	100	100
7th	14	14	100	100	100
8th	10	10	100	100	100
10th	16	16	100	100	100

Note: This table replicates Table 7 but uses the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Adjusted BIC to draw comparisons. Since Massachusetts's public data provides student-level item performance for all items, we are able to estimate a unidimensional measurement model for using all items. In addition, we estimate a version with only released assessment items to create a similar comparison to the bidimensional and testlet models which use only released assessment items. The lefthand columns of the table then report out the number of assessment combinations (grade-year-modality/session) and the number of models that fit the data without errors (e.g., negative variances or impossible correlations). The righthand columns of the table compare the fit statistics (using AIC, BIC, and Adjusted BIC) between the bidimensional and testlet models that fit the data to both unidimensional counterparts. The values represent the percentage of models that provide a superior fit across all fit statistics.

Appendix Table 6: Continuous Measurement Model Comparison Results Using Likelihood Ratio Tests

Bidimensional (Released Items)	**Assessments** (N)	**Converged Models (N)**	**Fit > Unidimensional (All Items)** (% of Converged Models)	**Fit > Unidimensional (Released Items)** (% of Converged Models)	**Fit > Unidimensional (Both All & Released)** (% of Converged Models)
MCAS 1.0	86	86	100	100	100
3rd	14	14	100	100	100
4th	14	14	100	100	100
5th	9	9	100	100	100
6th	9	9	100	100	100
7th	14	14	100	100	100
8th	10	10	100	100	100
10th	16	16	100	100	100

Note: This table replicates Table 7 but uses the Likelihood Ratio Test to draw comparisons. Since Massachusetts's public data provides student-level item performance for all items, we are able to estimate a unidimensional measurement model for using all items. In addition, we estimate a version with only released assessment items to create a similar comparison to the bidimensional and testlet models which use only released assessment items. The lefthand columns of the table then report out the number of assessment combinations (grade-year-modality/session) and the number of models that fit the data without errors (e.g., negative variances or impossible correlations). The righthand columns of the table compare the fit using the Likelihood Ratio Test between the bidimensional and testlet models that fit the data to both unidimensional counterparts. The values represent the percentage of models that provide a superior fit across all fit statistics.

Appendix Table 7: Alternative Outcomes Predictive Validity

	School's Percentage of Students Graduating High School Within 4-Years				School's Percentage of Students Enrolling in Postsecondary Institutions Within 16 Months			
	Unidimensional and Bidimensional				Unidimensional and Bidimensional			
ELA	3.38*** (0.332)				4.53*** (0.501)			
HOTS		3.22*** (0.322)		0.28 (0.258)		4.42*** (0.489)		1.56** (0.498)
LOTS			3.41*** (0.332)	3.15*** (0.368)			4.52*** (0.497)	3.06*** (0.569)

Note: percentage of school's postsecondary enrollment was created by using dividing the number of postsecondary enrollments within 16 months by the number of high schoolers that originally started in that school in 2009 based on a 4-year graduation rate. These results use the "Adjusted" populations (i.e., they do account for students transferring out of the school but not students transferring in). All Models include year fixed effects.

	School's Percentage of Students Graduating High School Within 5-Years				School's Percentage of Students Enrolling in Postsecondary Institutions Within 16 Months			
	Unidimensional and Bidimensional				Unidimensional and Bidimensional			
ELA	3.49*** (0.321)				6.51*** (0.398)			
HOTS		3.33*** (0.313)		0.32 (0.272)		6.26*** (0.391)		1.25** (0.440)
LOTS			3.53*** (0.321)	3.23*** (0.367)			6.54*** (0.396)	5.37*** (0.510)

Note: percentage of school's postsecondary enrollment was created by using dividing the number of postsecondary enrollments within 16 months by the number of high schoolers that originally started in that school in 2009 based on a 5-year graduation rate. These results use the "Unadjusted" populations (i.e., they account for students transferring in or out of the school). All Models include cohort fixed effects.

	School's Percentage of Students Graduating High School Within 5-Years				School's Percentage of Students Enrolling in Postsecondary Institutions Within 16 Months			
	Unidimensional and Bidimensional				Unidimensional and Bidimensional			
ELA	3.01*** (0.306)				4.53*** (0.502)			
HOTS		2.86*** (0.298)		0.28 (0.234)		4.42*** (0.490)		1.55** (0.498)
LOTS			3.03*** (0.306)	2.77*** (0.327)			4.52*** (0.498)	3.07*** (0.570)

Note: percentage of school's postsecondary enrollment was created by using dividing the number of postsecondary enrollments within 16 months by the number of high schoolers that originally started in that school in 2009 based on a 5-year graduation rate. These results use the "Adjusted" populations (i.e., they do account for students transferring out of the school but not students transferring in). All Models include cohort fixed effects.

Appendix Table 8: Alternative Outcomes Predictive Validity of 2-Year and 4-Year Postsecondary Institution Enrollment

	School's Percentage of Students Enrolling in 2-Year Postsecondary Institutions Within 16 Months				School's Percentage of Students Enrolling in 4-Year Postsecondary Institutions Within 16 Months			
	Unidimensional and Bidimensional				Unidimensional and Bidimensional			
ELA	-4.13*** (0.379)				8.66*** (0.543)			
HOTS		-3.91*** (0.370)		-0.16 (0.495)		8.33*** (0.536)		1.72* (0.716)
LOTS			-4.17*** (0.377)	-4.02*** (0.540)			8.69*** (0.540)	7.08*** (0.768)

Note: * p <0.05, ** p <0.01, and *** p<0.001. Percentage of school's postsecondary enrollment was created by dividing the number of postsecondary enrollments within 16 months by the number of high schoolers that originally started in that school in 2009. These results use the "Adjusted" populations (i.e., they do account for students transferring out of the school but not students transferring in). All Models include cohort fixed effects.

Appendix Table 9: Alternative Outcome - Original Postsecondary Enrollment Definition Predictive Validity

	School's % of High School Graduates Enrolling in Postsecondary			
	Unidimensional and Bidimensional			
ELA	4.81*** (0.313)			
HOTS		4.65*** (0.308)		1.23*** (0.372)
LOTS			4.81*** (0.310)	3.66*** (0.401)

Note: the outcome uses the original data available from Massachusetts - the percentage of high school graduates that enroll in postsecondary institutions within 16 months. All Models include cohort fixed effects.

Appendix Table 10: Bidimensional Measurement Models Predictive Validity with Data Aggregated to School-Level

4-Yr Unadjusted	School's Percentage of Students Graduating High School Within 4-Years				School's Percentage of Students Enrolling in Postsecondary Institutions Within 16 Months			
	Unidimensional and Bidimensional				Unidimensional and Bidimensional			
ELA	25.52*** (0.337)				32.30*** (0.312)			
HOTS		25.53*** (0.350)		1.91 (1.612)		32.68*** (0.322)		15.30*** (1.410)
LOTS			25.74*** (0.338)	23.89*** (1.594)			32.27*** (0.316)	17.58*** (1.390)
	School's Percentage of Students Enrolling in 2-Year Postsecondary Institutions Within 16 Months				School's Percentage of Students Enrolling in 4-Year Postsecondary Institutions Within 16 Months			
ELA	-7.94*** (0.256)				40.23*** (0.365)			
HOTS		-7.94*** (0.262)		-1.43 (1.162)		40.62*** (0.380)		16.75*** (1.651)
LOTS			-7.96*** (0.257)	-6.58*** (1.145)			40.22*** (0.370)	24.14*** (1.627)

Note: * p <0.05, ** p <0.01, and *** p<0.001. Percentage of school's postsecondary enrollment was created by dividing the number of postsecondary enrollments within 16 months by the number of high schoolers that originally started in that school in, for example, 2009. These results use the "Unadjusted" populations (i.e., they account for students transferring in or out of the school). These analyses start by aggregating individual students' scores on higher-order thinking, lower-order thinking, and ELA to the school-level. All Models include cohort fixed effects. Testlet models only use years 2005, 2010, 2013, and 2016. School's Percentage of Students Graduating High School Within 4-Years Mean Across All Years: 86.626%. School's Percentage of Students Enrolling in Postsecondary Institutions Within 16 Months Mean Across All Years: 66.696%. School's Percentage of Students Enrolling in 2-Year Postsecondary Institutions Within 16 Months Mean Across All Years: 18.101%. School's Percentage of Students Enrolling in 4-Year Postsecondary Institutions Within 16 Months Mean Across All Years: 48.587%.

Appendix Table 11: Bidimensional Measurement Models Predictive Validity by Year

	School's Percentage of Students Graduating High School Within 4-Years				School's Percentage of Students Enrolling in Postsecondary Institutions Within 16 Months			
Bidimensional Models	Single Predictors		Joint Predictors		Single Predictors		Joint Predictors	
	HOTS	LOTS	HOTS	LOTS	HOTS	LOTS	HOTS	LOTS
2003	5.04***	5.02***	2.89***	2.22**	7.92***	7.86***	4.98***	3.02***
	(0.486)	(0.484)	(0.734)	(0.704)	(0.509)	(0.502)	(0.968)	(0.901)
2004	4.75***	4.77***	1.53	3.25**	7.42***	7.38***	5.79***	1.65
	(0.456)	(0.454)	(1.255)	(1.229)	(0.468)	(0.462)	(1.556)	(1.460)
2005	4.59***	4.64***	1.42^	3.27***	7.16***	7.18***	3.29**	3.99***
	(0.438)	(0.431)	(0.812)	(0.751)	(0.476)	(0.463)	(1.079)	(0.983)
2006	4.20***	4.42***	-0.13	4.54***	6.48***	6.60***	2.08**	4.61***
	(0.407)	(0.412)	(0.649)	(0.708)	(0.421)	(0.420)	(0.810)	(0.813)
2007	4.18***	4.37***	-5.01***	9.31***	5.94***	6.14***	-4.59***	10.67***
	(0.425)	(0.430)	(1.183)	(1.294)	(0.406)	(0.408)	(1.335)	(1.404)
2008	4.03***	4.64***	-0.53	5.11***	5.71***	6.30***	0.40	5.94***
	(0.437)	(0.461)	(0.456)	(0.605)	(0.425)	(0.430)	(0.523)	(0.583)
2009	3.93***	4.18***	-0.06	4.24***	6.13***	6.29***	1.80**	4.60***
	(0.421)	(0.425)	(0.485)	(0.548)	(0.437)	(0.441)	(0.578)	(0.627)
2010	3.48***	3.84***	0.35	3.54***	5.47***	5.88***	1.22**	4.80***
	(0.364)	(0.384)	(0.349)	(0.466)	(0.391)	(0.404)	(0.474)	(0.545)
2011	3.72***	4.10***	0.21	3.91***	5.71***	6.11***	1.15*	5.08***
	(0.400)	(0.421)	(0.305)	(0.462)	(0.433)	(0.450)	(0.453)	(0.556)
2012	3.17***	3.55***	-0.08	3.63***	5.23***	5.70***	0.70	5.07***
	(0.346)	(0.371)	(0.289)	(0.447)	(0.397)	(0.410)	(0.483)	(0.561)
2013	3.31***	3.63***	0.38	3.30***	5.96***	6.32***	1.63***	4.87***
	(0.362)	(0.388)	(0.263)	(0.438)	(0.420)	(0.428)	(0.474)	(0.534)
2016	2.61***	2.77***	-0.14	2.90***	6.13***	6.47***	0.06	6.42***
	(0.244)	(0.256)	(0.480)	(0.548)	(0.405)	(0.406)	(0.701)	(0.730)
2017	2.40***	2.56***	0.47	2.14***	6.15***	6.75***	0.24	6.53***
	(0.243)	(0.246)	(0.442)	(0.461)	(0.444)	(0.431)	(0.635)	(0.631)

Note: * p <0.05, ** p <0.01, and *** p<0.001. Percentage of school's postsecondary enrollment was created by dividing the number of postsecondary enrollments within 16 months by the number of high schoolers that originally started in that school, for example, in 2009. These results use the "Unadjusted" populations (i.e., they account for students transferring in or out of the school). All Models use each individual cohort. School's Percentage of Students Graduating High School Within 4-Years Mean Across All Years: 86.626%. School's Percentage of Students Enrolling in Postsecondary Institutions Within 16 Months Mean Across All Years: 66.696%.

Appendix Table 12: All vs. Released Item Representativeness by Year and Grade

	All Items						Released Items					
		Subscale Reporting Category		Item Type				Subscale Reporting Category		Item Type		
Year	Total Items	Language	Reading and Literature	Multiple Choice	Open Response	Item Difficulty	Total Items	Language	Reading and Literature	Multiple Choice	Open Response	Item Difficulty
All Grades and 2009-2023	3977	19%	81%	90%	10%	-0.83	2142	18%	82%	90%	10%	-0.83
2009	282	14%	86%	91%	9%	-1.26	139	13%	87%	89%	11%	-1.30
2010	281	16%	84%	90%	10%	-1.11	140	16%	84%	89%	11%	-1.06
2011	281	18%	82%	90%	10%	-1.20	136	15%	85%	90%	10%	-1.20
2012	281	17%	83%	90%	10%	-1.13	147	16%	84%	89%	11%	-1.20
2013	281	18%	82%	90%	10%	-1.15	151	17%	83%	89%	11%	-1.15
2014	281	17%	83%	90%	10%	-1.06	138	14%	86%	89%	11%	-0.95
2015	281	17%	83%	90%	10%		146	16%	84%	89%	11%	
2016	281	16%	84%	90%	10%		138	15%	85%	90%	10%	
2017	340	24%	76%	88%	12%	-0.79	162	23%	77%	90%	10%	-0.94
2018	385	21%	79%	89%	11%	-0.54	233	18%	82%	89%	11%	-0.55
2019	350	23%	77%	89%	11%	-0.63	190	22%	78%	89%	11%	-0.51
2021	218	24%	76%	94%	6%	-0.46	137	25%	75%	94%	6%	-0.47
2022	217	25%	75%	94%	6%	-0.55	147	24%	76%	95%	5%	-0.53
2023	218	23%	77%	94%	6%	-0.37	138	22%	78%	94%	6%	-0.44
Grade 3 and All Years	907	21%	79%	92%	8%	-0.86	592	20%	80%	93%	7%	-0.86
Grade 4 and All Years	884	20%	80%	90%	10%	-0.89	588	20%	80%	90%	10%	-0.88
Grade 5 and All Years	684	19%	81%	90%	10%	-0.83	378	18%	82%	90%	10%	-0.88
Grade 6 and All Years	682	18%	82%	90%	10%	-0.77	386	16%	84%	90%	10%	-0.77
Grade 7 and All Years	885	16%	84%	90%	10%	-0.81	576	16%	84%	90%	10%	-0.80
Grade 8 and All Years	724	18%	82%	90%	10%	-0.91	408	16%	84%	90%	10%	-0.92

Note: All Grade 10 Items Are Released. The table reports out the total of "All Items" followed the percentage that belong to each subscale reporting category and item type/format. It also reports out the average item difficulty for all items. The process repeats but for "Released Items". In the 2015 and 2016 school years, the public data do not allow us to estimate item difficulties, so those are left blank. This table breakdowns the item characteristics for each year and grade.

Appendix Table 13: Regressions of Item Characteristics on Released by Year

	Regression of Language on Released				Regression of Multiple Choice on Released				Regression of Item Difficulty on Released			
Year	Estimate	Standard Error	Test Statistic	P Value	Estimate	Standard Error	Test Statistic	P Value	Estimate	Standard Error	Test Statistic	P Value
All Grades and 2009-2023	-0.01	0.01	-0.72	0.47	0.00	0.01	0.08	0.94	-0.01	0.03	-0.38	0.70
2009	-0.01	0.05	-0.31	0.76	-0.03	0.04	-0.80	0.42	-0.08	0.10	-0.85	0.40
2010	0.02	0.05	0.40	0.69	-0.01	0.04	-0.26	0.80	0.11	0.09	1.29	0.20
2011	-0.06	0.05	-1.17	0.24	0.02	0.04	0.42	0.68	0.00	0.09	-0.04	0.97
2012	-0.02	0.05	-0.39	0.70	-0.02	0.04	-0.39	0.70	-0.15	0.09	-1.64	0.10
2013	0.01	0.05	0.10	0.92	-0.03	0.04	-0.64	0.53	-0.02	0.09	-0.21	0.83
2014	-0.05	0.05	-0.89	0.37	-0.01	0.04	-0.34	0.73	0.23	0.09	2.50	0.01
2015	0.00	0.05	-0.07	0.95	-0.02	0.04	-0.38	0.71				
2016	-0.02	0.05	-0.31	0.76	0.00	0.04	0.07	0.94				
2017	0.03	0.05	0.66	0.51	0.04	0.04	0.94	0.35	-0.32	0.13	-2.58	0.01
2018	-0.03	0.05	-0.51	0.61	0.00	0.04	0.03	0.98	0.00	0.09	0.00	1.00
2019	-0.01	0.06	-0.24	0.81	0.00	0.04	0.03	0.98	0.19	0.13	1.49	0.14
2021	0.04	0.07	0.66	0.51	0.02	0.04	0.50	0.62	0.01	0.20	0.03	0.98
2022	-0.03	0.07	-0.49	0.62	0.03	0.04	0.91	0.36	0.07	0.11	0.68	0.50
2023	-0.02	0.06	-0.30	0.76	0.02	0.04	0.53	0.60	-0.21	0.10	-2.09	0.04

Note: All Grade 10 Items Are Released, so they are not included. Linear regressions of the subscale (Language), item format (Multiple Choice), or item difficulty on whether the item is released are estimated. All regressions include grade-year-modality fixed effects to ensure assessments are compared to themselves. The results of the regression are reported out with the coefficient (Estimate), standard error, test statistic, and p-value. This table repeats these regressions for each individual year.

Appendix Table 14: Regressions of Item Characteristics on Released by Grade

	Regression of Language on Released				Regression of Multiple Choice on Released				Regression of Item Difficulty on Released			
Year	Estimate	Standard Error	Test Statistic	P Value	Estimate	Standard Error	Test Statistic	P Value	Estimate	Standard Error	Test Statistic	P Value
All Grades and 2009-2023	-0.01	0.01	-0.72	0.47	0.00	0.01	0.08	0.94	-0.01	0.03	-0.38	0.70
Grade 3 and 2009-2023	-0.05	0.04	-1.49	0.14	0.01	0.03	0.51	0.61	0.00	0.08	0.02	0.99
Grade 4 and 2009-2023	0.03	0.04	0.79	0.43	0.01	0.03	0.34	0.74	0.02	0.07	0.26	0.79
Grade 5 and 2009-2023	-0.01	0.04	-0.29	0.77	-0.01	0.03	-0.28	0.78	-0.12	0.07	-1.58	0.11
Grade 6 and 2009-2023	-0.02	0.03	-0.60	0.55	0.00	0.03	0.03	0.98	-0.03	0.07	-0.43	0.67
Grade 7 and 2009-2023	0.01	0.03	0.28	0.78	0.00	0.03	0.02	0.98	0.05	0.06	0.75	0.46
Grade 8 and 2009-2023	-0.01	0.03	-0.40	0.69	-0.01	0.03	-0.46	0.65	0.01	0.07	0.10	0.92

Note: All Grade 10 Items Are Released, so they are not included. Linear regressions of the subscale (Language), item format (Multiple Choice), or item difficulty on whether the item is released are estimated. All regressions include grade-year-modality fixed effects to ensure assessments are compared to themselves. The results of the regression are reported out with the coefficient (Estimate), standard error, test statistic, and p-value. This table repeats these regressions for each individual grade.

Appendix Table 15: Logistic Regressions of Item Characteristics on Released by Year and Grade

	Logistic Regression of Language on Released				Logistic Regression of Multiple Choice on Released			
Year	Estimate	Standard Error	Test Statistic	P Value	Estimate	Standard Error	Test Statistic	P Value
All Grades and 2009-2023	-0.06	0.09	-0.73	0.47	0.01	0.12	0.08	0.94
2009	-0.12	0.38	-0.31	0.76	-0.36	0.45	-0.81	0.42
2010	0.15	0.36	0.41	0.68	-0.11	0.43	-0.26	0.80
2011	-0.42	0.36	-1.18	0.24	0.19	0.44	0.42	0.67
2012	-0.14	0.35	-0.39	0.70	-0.17	0.42	-0.39	0.69
2013	0.04	0.34	0.11	0.92	-0.28	0.43	-0.64	0.52
2014	-0.32	0.35	-0.90	0.37	-0.15	0.43	-0.35	0.73
2015	-0.02	0.34	-0.07	0.95	-0.16	0.43	-0.38	0.70
2016	-0.11	0.35	-0.31	0.76	0.03	0.43	0.07	0.94
2017	0.18	0.27	0.67	0.50	0.36	0.38	0.96	0.34
2018	-0.15	0.29	-0.52	0.60	0.01	0.38	0.03	0.98
2019	-0.07	0.29	-0.25	0.81	0.01	0.38	0.03	0.98
2021	0.24	0.35	0.67	0.50	0.31	0.61	0.51	0.61
2022	-0.17	0.35	-0.50	0.62	0.55	0.60	0.92	0.36
2023	-0.11	0.35	-0.31	0.76	0.32	0.60	0.54	0.59
Grade 3 and 2009-2023	-0.31	0.21	-1.51	0.13	0.14	0.28	0.51	0.61
Grade 4 and 2009-2023	0.17	0.21	0.80	0.42	0.10	0.29	0.34	0.73
Grade 5 and 2009-2023	-0.07	0.22	-0.30	0.77	-0.08	0.29	-0.28	0.78
Grade 6 and 2009-2023	-0.14	0.22	-0.61	0.54	0.01	0.29	0.03	0.98
Grade 7 and 2009-2023	0.07	0.23	0.28	0.78	0.01	0.29	0.02	0.98
Grade 8 and 2009-2023	-0.09	0.22	-0.41	0.68	-0.14	0.29	-0.47	0.64

Note: All Grade 10 Items Are Released, so they are not included. Logistic regressions of the subscale (Language) and item format (Multiple Choice) on whether the item is released are estimated. All regressions include grade-year-modality fixed effects to ensure assessments are compared to themselves. The results of the regression are reported out with the coefficient (Estimate), standard error, test statistic, and p-value. This table repeats these regressions for each individual year and grade.

Appendix Table 16: Item Characteristics by Test Regime

	Item Counts		Item Subscale				Item Format			
Item Characteristic Overlap	HOTS Count	LOTS Count	HOTS Language	HOTS Reading and Literature	LOTS Language	LOTS Reading and Literature	HOTS Multiple Choice	HOTS Open-Response	LOTS Multiple Choice	LOTS Open-Response
All Grades & Years	649	2710	11%	89%	19%	81%	53%	47%	100%	0%
MCAS 1.0	440	2188	2%	98%	18%	82%	44%	56%	100%	0%
3rd	35	402	3%	97%	20%	80%	26%	74%	100%	0%
4th	65	356	2%	98%	18%	82%	37%	63%	100%	0%
5th	39	183	0%	100%	18%	82%	41%	59%	100%	0%
6th	39	192	0%	100%	17%	83%	38%	62%	100%	0%
7th	74	348	0%	100%	17%	83%	42%	58%	100%	0%
8th	45	210	2%	98%	18%	82%	40%	60%	100%	0%
10th	143	497	3%	97%	17%	83%	55%	45%	100%	0%
MCAS 2.0	209	522	30%	70%	21%	79%	72%	28%	100%	0%
3rd	14	81	36%	64%	23%	77%	50%	50%	100%	0%
4th	23	89	39%	61%	28%	72%	61%	39%	100%	0%
5th	29	69	34%	66%	20%	80%	72%	28%	100%	0%
6th	31	74	29%	71%	16%	84%	71%	29%	100%	0%
7th	32	67	28%	72%	22%	78%	75%	25%	100%	0%
8th	35	67	34%	66%	12%	88%	74%	26%	100%	0%
10th	45	75	18%	82%	23%	77%	82%	18%	100%	0%

Note: The table shows only released assessment items (since they could be coded). Items are broken down into whether they were coded as HOTS or LOTS, and the subsequent columns provide the percentage of those HOTS or LOTS items that fall under each subscale or format. These are separated by testing regime.

Appendix Table 17: Breakdown of HOTS or LOTS in ELA Standards

	All ELA Standards				All ELA Standards with Five or More Released Items			
ELA Standards	Total Standards (N)	Only LOTS Items Standards (%)	Only HOTS Items Standards (%)	Both LOTS and HOTS Items (%)	Total Standards (N)	Only LOTS Items Standards (%)	Only HOTS Items Standards (%)	Both LOTS and HOTS Items (%)
All Grades & Years	288	35%	4%	61%	187	25%	0%	75%
MCAS 1.0	135	30%	4%	67%	108	22%	0%	78%
3rd	17	41%	0%	59%	15	33%	0%	67%
4th	18	33%	0%	67%	14	14%	0%	86%
5th	15	27%	7%	67%	13	23%	0%	77%
6th	17	29%	6%	65%	13	15%	0%	85%
7th	18	28%	0%	72%	13	23%	0%	77%
8th	17	24%	0%	76%	13	23%	0%	77%
10th	33	27%	9%	64%	27	22%	0%	78%
MCAS 2.0	153	39%	4%	57%	79	28%	0%	72%
3rd	22	55%	0%	45%	11	45%	0%	55%
4th	22	50%	5%	45%	11	45%	0%	55%
5th	20	40%	10%	50%	10	20%	0%	80%
6th	21	29%	0%	71%	14	29%	0%	71%
7th	23	30%	4%	65%	11	9%	0%	91%
8th	23	30%	9%	61%	9	11%	0%	89%
10th	22	41%	0%	59%	13	31%	0%	69%

Note: The table documents the number of standards by test regime and grade. The subsequent columns report the percentage of those standards that have only items coded as LOTS, HOTS, or both. The second panel repeats this but limits the ELA standards to only those that feature at least five or more released items (to avoid the problem of few items driving conclusions).

Appendix Table 21: Bidimensional Scalar Measurement Invariance Met

Bidimensional (Released Items) Scalar Measurement Invariance Met Change in CFI (0.01), RMSEA (0.015), and SRMR (0.030 for Metric & 0.015 for Scalar) Criteria			
Years	Converged Models (N)	SES (%)	Race (%)
Entire Massachusetts Sample	77	95%	97%
3rd	10	90%	80%
4th	12	92%	100%
5th	8	88%	100%
6th	8	100%	100%
7th	14	100%	100%
8th	9	89%	100%
10th	16	100%	100%
2001	5	80%	100%
2002	4	100%	100%
2003	3	100%	100%
2004	4	100%	100%
2005	4	100%	100%
2006	7	100%	86%
2007	6	100%	100%
2008	7	100%	100%
2009	7	86%	86%
2010	5	100%	100%
2011	4	100%	75%
2012	7	57%	100%
2013	7	100%	100%
2014	6	100%	100%
2017	1	100%	100%
2018	1	100%	100%

Note: Comparative Fit Index (CFI), Root Mean Squared Error of Approximation (RMSEA), and Standardized Root Mean Squared Residuals (SRMR). This table summarizes the extent scalar measurement invariance is met using Chen's 2007 criteria among all converging higher-order and lower-order thinking bidimensional measurement models for socioeconomic status (economically disadvantaged versus non-economically disadvantaged) and race (Asian, Black, Hispanic, and White). Other racial groups are excluded from the measurement invariance tests and analyses because the sample sizes are too small. The subsequent rows examine the extent measurement invariance is met for individual grades and years.

Appendix Figure X: Example of HOTS and LOTS with Low and High Item Difficulty

	LOTS	HOTS
Low Difficulty	Based on the passage, what causes Tomas to fall off the boat? Ⓐ A whale knocks Tomas out of the boat. Ⓑ The boat rocks too hard and makes Tomas fall. Ⓒ Tomas leans too far over the side of the boat to touch the whale. Ⓓ Dad accidentally shoves Tomas while running to the side of the boat.	What is the central message of the passage? Ⓐ It is important to always be prepared. Ⓑ People learn best by trying new things. Ⓒ Making a family member happy has its rewards. Ⓓ Things can sometimes turn out differently than we expect.
	Item Difficulty: -0.45	**Item Difficulty: -0.53**
High Difficulty	Read the dictionary entry in the box. **head:** *v.* **1.** to lead something **2.** to go in a certain direction **3.** to give a title to **4.** to place at the beginning of Based on the passage, which meaning of the word ***head*** is used in paragraph 1? Ⓐ meaning 1 Ⓑ meaning 2 Ⓒ meaning 3 Ⓓ meaning 4	Based on *A Vacation in Ruins*, write an essay that explains how Marisol's feelings change throughout the passage. Be sure to use information from the passage to develop your essay.
	Item Difficulty: 0.47	**Item Difficulty: 1.09**

Note: These items are sourced from the Massachusetts Department of Elementary and Secondary Education's publicly available 2023 Grade 3 assessment, which permits reproduction for non-commercial purposes. The items are based on a passage from *A Vacation in Ruins* by Precious McKenzie, which is not shown due to its copyright. Item difficulties are generated using Item Response Theory Three-Parameter Logistic and Graded Response Models with all items (including those without text released). The rationale for each item's categorization as HOTS or LOTS are as follows: (LOTS, Low Difficulty) item asks students to summarize an event in the text; (LOTS, High Difficulty) item asks students to identify the meaning of a word in context; (HOTS, Low Difficulty) item asks students to identify/make inferences about explicit or implicit themes; and (HOTS, High Difficulty) item asks students to explain, generalize, or connect ideas using supporting evidence.