MARYLAND STATE DEPARTMENT OF EDUCATION
EQUITY AND EXCELLENCE

ISR Language Evaluation:

Gathering Stakeholder Feedback to Make Community-Informed Program Improvements

By
Susan Lyons, PhD
Fiona Hinds, EdD
Sanford Student
Hannah Denker

cognia™

Lyons ASSESSMENT CONSULTING

MAY 2022

We are grateful for the extensive collaboration, leadership, and thought partnership of Brianna Creed and Paul Katula at the Maryland State Department of Education, who have been instrumental in the design and implementation of this evaluation effort.

Suggested Citation: Lyons, S., Hinds, B. F., Student, S., & Denker, H. (2022). *ISR Language Evaluation: Gathering Stakeholder Feedback to Make Community-Informed Program Improvements.* Cognia and Lyons Assessment Consulting.

Table of Contents

Executive summary and recommendations

Prior research has demonstrated that the language of the achievement level labels on Individual Student Reports (ISR) can affect feelings of encouragement and motivation in students, particularly for those students scoring in the lowest level who are disproportionately students of color (O'Donnell, 2020). Given these prior findings, the Maryland State Department of Education (MSDE) has partnered with Cognia and Lyons Assessment Consulting to conduct an extensive stakeholder feedback initiative to make community-informed improvements to the language of the achievement level labels used in the Maryland Comprehensive Assessment Program. This effort was conducted with the intention of aligning Maryland's achievement level language with the prior literature related to the relationship between the labels and student perceptions of their future potential, and to more clearly express MSDE's belief in the potential for every student to excel. This evaluation explored the extent to which students, caregivers, and educators perceived a variety of achievement level labels as encouraging. As a result, the recommendations in this report represent a systematic way to value the input of the state's most important stakeholders in making evidence-based program improvements.

Our evaluation was based upon qualitative and quantitative analysis of data from a large-scale survey including both selected- and open-response items, as well as caregiver focus groups. By working directly with a diverse set of Maryland school districts, we were able to collect responses from over 4,000 Maryland students, more than 1,700 Maryland caregivers, and more than 500 Maryland educators across the state. We surveyed respondents on their preferences among current and alternate language for each of Maryland's four achievement levels, with four options at Levels 1–2 and three at Levels 3–4.

The findings from our quantitative and qualitative analyses largely converged. We recommend the following revised labels for Maryland's ELA, math, science and social studies achievement levels:

- Level 1: Beginning (in need of support)
- Level 2: Approaching Expectations
- Level 3: Meeting Expectations
- Level 4: Exceeding Expectations

We also strongly recommend developing additional interpretive supports for caregivers and educators that provide information related to the intended meaning of the achievement levels and any appropriate actions that may be warranted due to the student's score. We suggest including a description of each of the performance levels on the ISRs, as well as adding or enhancing the available score report interpretations guides.

This is especially important for those students scoring below proficiency. For example, for students scoring within the lowest achievement level, additional interpretive language for caregivers could be the following: "Your child is scoring in the lowest level measured by the state assessment and is demonstrating *below grade level* knowledge and skills. *Additional and immediate support is needed* to ensure your child is on track for college and career readiness." Supporting language like this can serve as a call to action for parents and educators to advocate for or provide immediate and needed educational interventions for students who are at risk of graduating without the necessary knowledge and skills for meaningfully engaging in their post-secondary plans. While the state assessment is not able to provide detailed information at a granular size that would be useful for informing the exact type of educational intervention needed (e.g., where gaps or misunderstandings may exist), the state assessment can and should be a strong signal to parents and educators about the degree to which the student's current educational program is adequately meeting their needs for reaching the state's grade level expectations. By providing this additional context, we expect that Maryland can simultaneously use the kinds of labels that students find most encouraging, which was the highest priority of this evaluation, and provide the context and urgency that caregivers and educators appreciate.

Introduction

This report describes the result of a collaboration between the Maryland Department of Education (MSDE), Cognia, and Lyons Assessment Consulting to evaluate the impact that the language the state uses to communicate about student assessment performance has on students' self-perceptions of academic potential. Through this effort, we sought to revise the language of the achievement levels labels to terminology that Maryland students, caregivers, and teachers would find more encouraging, with special attention to the students who are performing below grade level or were likely negatively affected by current ISR language.

Rationale for evaluation

Research has confirmed that the belief that one's intellectual abilities can be developed, referred to as a growth mindset (Dweck, 2008), is related to improved performance (Dweck, 2000; Claro, Paunesku & Dweck, 2016; Yeager et al., 2019). Growth mindset is associated with feelings of self-efficacy and student motivation—important factors in student learning (National Academies of Sciences, Engineering, and Medicine, 2018). Learning about the malleability of intelligence has been shown to be particularly powerful in improving outcomes for racial and ethnic minority students (Aronson, Fried, & Good, 2002; Blackwell, Trzesniewkski & Dweck, 2007; Broda et al., 2018). While the empirical literature connecting state assessment score reporting to student self-perceptions and student achievement is sparse, the little evidence that exists suggests that more growth-oriented labeling is associated with improvements in academic performance and an increased likelihood to attend college among low-income and minoritized student groups (Papay et al., 2016).

Maryland's current achievement level descriptors were not designed to reflect a growth mindset. To align with the literature and more effectively communicate MSDE's belief in the potential for every student to excel, this evaluation explored the extent to which stakeholders in MSDE's score reporting practices perceived a variety of potential achievement level labels as encouraging related to student potential for future achievement.

Theoretical framework

The research on which we could draw to inform our alternate labels is sparse; there is, to our knowledge, only one empirical quantitative study of score report achievement level labeling with direct relevance to this study (O'Donnell, 2020); one other study does exist (Burt & Stapleton, 2010), but O'Donnell's work is more comprehensive, is more focused on the types of affective responses in which we are interested, and includes caregivers, students, and teachers, not just teachers. Consequently, we primarily drew upon O'Donnell's work.

Currently, of the four performance levels defined for Maryland students who take end-of-year assessments in math and reading, the two levels associated with the lowest scale scores are called "minimal understanding" and "partial understanding." Prior research has shown that these labels are neither encouraging nor clear (O'Donnell, 2020). With these criteria in mind, we suggested potential alternatives, which are outlined below. We also suggested alternatives to the current language for levels three and four, "satisfactory understanding" and "extensive understanding."

The first suggested revision centered on expectations for students; we suggested replacing the four current level names with "not yet meeting expectations," "partially meeting expectations," "meeting expectations," and "exceeding expectations." Prior research indicates that these labels balance encouragement with clarity in a study where caregivers, teachers, and students were asked to compare a variety of labels for achievement levels on score reports (O'Donnell, 2020). Students found the label "not yet meeting expectations" for the lowest achievement level especially encouraging.

Another potential option was to use performance level labels that are oriented toward actionability–that is, the labels describe how a student at that level can meet the standards to which the report is written. As found in O'Donnell (2020), the label "in need of support" for the lowest achievement level was found to be both very encouraging and very clear. It was also noted that this label reinforced students' belief in their own capability. However, this label did not have an obvious counterpart for other performance levels, nor was it clear that it should apply only to the lowest performance level when students at the second lowest level presumably are also in need of support to meet standards, albeit to a lesser extent. We suggested the following: "in need of support," "approaching expectations," "meeting expectations," and "exceeding expectations."

Finally, the third set of descriptors was oriented toward framing students as progressing in their learning, whether their current score places them at the low end or the high end of the scale. These descriptors were "beginning," "developing," "on-track" and "advancing." Although these labels are not all explored in depth in O'Donnell (2020), we believed that it was important to explore a set of labels that decenters the expectations of MSDE and focuses more explicitly on the students' learning progressions.

Evaluation questions

The MSDE sought feedback from students, caregivers, and educators in order to make community-informed improvements to the current ISRs. We engaged students in responding to surveys to gain insight into the following questions:

1. Which growth mindset-oriented revision to Maryland's Performance Levels (PL) is the most encouraging to elementary, middle, and high school students?
2. Is this consistent across different racial/ethnic subgroups and student ages? If not, what differences do we find?

The engagement with caregivers focused on their perspectives on their child's learning:

3. Which growth mindset-oriented revision to Maryland's Performance Levels (PL) is the most encouraging to the caregivers of elementary, middle, and high school students?

And finally, we engaged with teachers to better understand a related question from the perspective of educators:

4. Which growth mindset-oriented revision to Maryland's Performance Levels (PL) is the most encouraging to K-12 educators?

Analytic methods

Participant recruitment

Our aim with this evaluation was to understand how changes to ISR wording would be perceived within the state of Maryland. Therefore, the goals for our stakeholder feedback collection were as follows:

- Data are fairly representative of Maryland as a whole
- Sample sizes are sufficient to support subgroup-level analyses, where required
- The evaluation design targets inferences about the ordering of PL labels
- Data collection is feasible

All Maryland school districts were invited to participate in this study via emails to each of the state's 24 district superintendents. This was followed by further outreach to specific districts that, combined, produced a sample largely representative of Maryland as a whole. Ultimately, the following seven districts participated in the study:

- Calvert County
- Caroline County
- Carroll County Public Schools
- Charles County
- Prince George's County Public Schools
- Somerset County Public Schools
- Worcester County Public Schools

In collaboration with MSDE, each district's personnel arrived at a plan for recruiting educators, students, and caregivers to participate in the study. These plans mainly involved emails to caregivers and setting aside time during the school day for students to take a survey. Districts were asked to only have students in grades 5–12 take the survey to ensure both sufficient reading level and prior experience with ISRs. Caregivers of all current district students were invited to take the survey, as were all teachers in participating schools. The data collection details are described in the following section.

Quantitative data collection

For each of the three stakeholder groups, we administered a survey consisting primarily of selected-response, forced-choice comparison items. The survey presented respondents with an example of an ISR section where achievement level labels are used to describe student performance. This was followed by 18 items asking the respondent which of two wording options they found more encouraging, using the following questions:

- Which of these descriptions would you find more encouraging to describe your performance on the state test? (Students)
- Which of these descriptions would you find more encouraging to describe a student's performance on the state test? (Educators)
- Which of these descriptions would you find more encouraging to describe your child's performance on the state test? (Caregivers)

These were followed by two open-ended items asking respondents to explain their choice between (1) the current level 1 wording ("minimal understanding") and one of the three alternatives presented in the survey and (2) the current level 2 wording ("partial understanding") and one of the three alternatives presented in the survey. The alternative language choice was presented at random from the three available options.

Finally, we asked participants to respond to demographic questions about gender, race/ethnicity, language spoken at home, and grade (of the student, of one's children, or that one teaches).

The details of each survey can be found in Appendices A–C starting on page 22. Surveys were distributed via direct links and QR codes and were administered using the SurveyMonkey platform.

Quantitative data analysis

Each stakeholder group's responses to the forced-choice survey items were analyzed to answer the following questions:

1. What is the overall order of this group's label preferences for each achievement level?
2. Are the group's preferences statistically distinguishable from the current ISR language for that level?

Additionally, for students, we asked:

3. Do preferences differ for specific student groups?

The first question can be answered for a given performance level by taking the simple sum of the number of times each option was chosen. Because we asked respondents all possible combinations of two wording choices at each level, these sums express the extent to which each language choice was preferred overall, and their order corresponds to respondents' overall order preference.

After establishing the preferred order for each group at each level, we addressed the second question using a hypothesis test. We began by identifying all wording choices at each level that were preferred to the current operational wording at that level. Then, using responses to the item that asked to directly compare the current wording with each preferred wording, we tested the null hypothesis that the proportion of people favoring the preferred choice in the underlying population

of which the sample is representative is 0.5 or less (making this a one-sided test). A *p*-value of 0.01* was taken as support for rejecting this null hypothesis, leading to the conclusion that the proportion of respondents preferring the alternative language was truly greater than 0.5. Rejecting the null hypothesis corresponds to the conclusion that the results we found are very unlikely to have arisen by chance. All analysis was conducted using R software (R Core Team, 2020). We used the tidyverse (Wickham et al., 2019) for data cleaning/summarizing and the prop.test built-in function for hypothesis testing.

We addressed the third question by replicating our analyses for student groups according to their grade and race/ethnicity.

Qualitative data collection

We used two sources of data for the qualitative analysis. First, we collected unique, open-ended survey responses from the participating educators, students, and caregivers describing their rationale for choosing their Level 1 and Level 2 language preferences. The first question asked, "In this survey, you were asked to choose between the terms 'minimal understanding' and [one of the three Level 1 revisions]. Please explain why you chose the wording that you did." The second question asked, "In this survey, you were asked to choose between the terms 'partial understanding' and [one of the three Level 2 revisions]. Please explain why you chose the wording that you did." In both cases, stakeholders were asked to compare the current language at a given level to one of the three revisions.

Next, we invited all participating caregivers to attend focus groups to share more information about their preferences. We held nine focus groups with 33 caregivers via Zoom during March 2022, which enabled us to collect data through a closed captioning transcript, a chat feature, and notes captured by one of the Lyons Assessment Consulting team members. Each semi-structured focus group lasted 45 minutes and included 20 minutes for caregivers to engage with discussion prompts that encouraged them to add information about why they selected their preferred language over the other options.

Qualitative data processing

The response variables from the survey data were first cleaned by limiting the data to only observations where the respondent completed the open-response item for both Level 1 and Level 2. From this sample of responses, we randomly selected 100 survey responses from each of the caregiver, educator, and student groups. This resulted in 600 excerpts across stakeholders to include in our qualitative analysis of survey responses. These 600 excerpts were imported into a qualitative analysis software package, Dedoose.

The caregiver focus group transcript, chat, and notes were reconciled to produce 76 excerpts that were included in our qualitative analysis of the Level 1 and Level 2 language. A speech utterance was counted as an individual excerpt if it occurred after another caregiver spoke. An utterance that included more than one sentence was added to the existing excerpt until the speaker changed. This also applied to utterances from the chat feature of Zoom. The 76 excerpts from the focus groups were also brought into Dedoose. These excerpts were analyzed separately from the survey data at first to ensure that the two sources of data did not conflict with each other. If they did, we would

* Due to the many hypothesis tests, we opted for a threshold of 0.01 to help account for the inflated type-1 error rate.

have relied more heavily on the survey responses as they were potentially more representative of the population of caregivers of interest. However, we found relatively similar results across focus groups and caregiver surveys so the data were analyzed together for the final results.

Qualitative data analysis

Each excerpt was coded in Dedoose for the word choice the respondent saw for Level 1 (i.e., minimal understanding v. beginning v. in need of support v. not yet meeting expectations) or saw for Level 2 (i.e., partial understanding v. approaching expectations v. developing v. partially meeting expectations). We used codes to identify excerpts where respondents specifically mentioned their preference. We also coded for respondent type (i.e., caregiver, educator, or student). Finally, we coded each excerpt for the respondent's main rationale for their choice (i.e., clarity and/or encouragement). For example, an educator excerpt that reads "I chose Minimal Understanding because it was more clear and Beginning seemed vague" would receive the following codes: Level 1; Minimal Understanding v. Beginning; Minimal Understanding; Educator; Clarity.

We identified subthemes, using inductive coding, within each of the two main themes listed above, letting subthemes emerge from the data. The first subthemes to emerge included Makes More Clear v. Unclear and Encouraging v. Discouraging. We also coded excerpts that did not indicate a clear preference for language (No Preference) and that provided a different rationale for a choice outside of encouragement or clarity (Other). We then reiterated the inductive coding process to refine these initial four subthemes into more specific reasons for their choices. This included downloading the excerpts by subtheme into Excel, reading through the excerpts, and identifying one or more phrases that gave insight into the respondents' reasoning. These key phrases were then color-coded according to the insight they provided. Like-color phrases were grouped together and reread to let the more specific subtheme arise from the language of the respondents themselves. Additionally, we identified words that were being used frequently within each subtheme (e.g., positive, harsh) and key quotes that capture common ideas that many respondents were sharing. Thus, we were able to identify subthemes within the stakeholder group and the main theme.

Quantitative results

Participants

Students

We received 4,184 total student responses. From this full sample, we filtered out any student who listed a grade below 5, multiple grades, or said their grade was not listed. This resulted in a final sample of 3,909 students. Below, we present the breakdown of the sample by grade, gender, and race/ethnicity.

The sample was fairly balanced by grade level, with at least 270 respondents from each grade 5–12.

Table 1. Student sample counts by grade

Grade	N
5	318
6	802
7	747
8	677
9	401
10	357
11	337
12	270

Sample sizes were more than sufficient to support comparisons of preferences by grade level groups, shown below.

The sample was well-balanced on gender, with a nearly 50/50 split between male and female respondents, as well as small but non-negligible groups of students who listed another gender or elected not to respond to the question.

Table 2. Student sample counts by gender

Gender	N
Female	1,799
Male	1,733
Other/not listed	171
Did not respond	206

The sample was about 41% white, about 27% Black, 11% multiple race/ethnicity, 7% Hispanic/Latino, and less than 5% of each additional group. About 8% of respondents did not answer this question.

Table 3. Student sample counts by race/ethnicity

Race/ethnicity	N
American Indian/Alaska Native	46
Asian	66
Black/African American	1,035
Hawaiian Native/Pacific Islander	7
Hispanic/Latino	269
White	1,604
Other race/ethnicity	135
Multiple race/ethnicity	447
Did not respond	300

Caregivers

We received 1,773 total caregiver responses. Of the respondents, 63 either did not list a child's grade or responded that their child's grade was not listed. We removed these respondents from the sample, leading to a final sample size of 1,710.

The breakdown of caregivers' children's grades is below.

Table 4. Caregiver sample counts by child grade

Grade	N
K	96
1	85
2	75
3	66
4	58
5	63
6	52
7	48
8	72
9	64
10	68
11	78
12	47
Multiple grades	838

About half of the respondents have children in more than one grade. Among the remaining respondents, the distribution across grades is pretty even, though the youngest grades are slightly overrepresented.

Turning to gender, the caregiver sample is majority-female.

Table 5. Caregiver sample counts by gender

Gender	N
Female	1,256
Male	209
Other/not listed	21
Did not respond	224

For race/ethnicity, the caregiver sample is majority-white.

Table 6. Caregiver sample counts by race/ethnicity

Race/ethnicity	N
American Indian/Alaska Native	4
Asian	21
Black/African American	93
Hawaiian Native/Pacific Islander	1
Hispanic/Latino	30
White	1,193
Other race/ethnicity	31
Multiple race/ethnicity	95
Did not respond	242

Educators

We received 542 total educator responses. Of the respondents, 77 either did not list a grade or responded that their taught grade was not listed. However, unlike for educators and students, we did not remove these responses. We presumed that they came from administrators, such as principals, or possibly student-facing staff, such as special education support personnel, and we considered their input valuable.

The breakdown of educators' taught grades is below. A majority of the sample teaches students in multiple grades.

Table 7. Caregiver sample counts by child grade

Grade	N
K	17
1	6
2	11
3	12
4	22
5	21
6	17
7	26
8	13
9	12
10	13
11	7
12	8
Multiple grades	280
No response or not listed	77

Turning to gender, the educator sample is majority-female.

Table 8. Caregiver sample counts by gender

Gender	N
Female	368
Male	63
Other/not listed	6
Did not respond	105

For race/ethnicity, the sample is majority-white.

Table 9. Caregiver sample counts by race/ethnicity

Race/ethnicity	N
American Indian/Alaska Native	1
Asian	6
Black/African American	46
Hawaiian Native/Pacific Islander	0
Hispanic/Latino	5
White	348
Other race/ethnicity	5
Multiple race/ethnicity	15
Did not respond	116

Summary of Findings

Overall, each group expressed a significant preference for language other than the current ISR achievement level labels at every achievement level. At Levels 3 and 4, there was general agreement that "meeting expectations" and "exceeding expectations" are more encouraging than the current language; at Level 3, the term "on-track" was also well-received. At Level 2, "approaching expectations" and "developing" were both universally preferred to the current wording, with "approaching expectations" the overall favorite. The clearest disagreement between stakeholder groups was at Level 1, where students' strong preference for "beginning" contrasted with caregivers' preference for "in need of support." Still, across the three groups, "beginning" does appear to be the overall favorite, as even caregivers favored it in a direct comparison to the current language.

Findings by stakeholder group

Students

The number of total times that each label was preferred by a student respondent is listed in Table 10. In this table, the current wording used on ISRs is italicized. For labels with an N preferred value above that of the current wording, an asterisk indicates that in a direct comparison item, this label was preferred to the current wording at a statistically significant ($p < 0.01$) level. Note that for every hypothesis test, the p-value was in fact 0.001 or below.

Table 10. Overall student preference counts

Achievement level	Label	N preferred
Level 1	Beginning	6,916*
	Minimal understanding	6,222
	In need of support	5,191
	Not yet meeting expectations	4,911
Level 2	Approaching expectations	7,238*
	Developing	6,324*
	Partial understanding	4,968
	Partially meeting expectations	4,704
Level 3	On-track	4,187*
	Meeting expectations	4,098*
	Satisfactory understanding	3,350
Level 4	Exceeding expectations	4,536*
	Advancing	3,819*
	Extensive understanding	3,267

*Significantly different from reference language ($p < 0.001$)

Subgroup sensitivity checks. To ensure that our overall findings for students were not obscuring the preferences of historically marginalized groups or differences by students' age, we ran the same analyses outlined above for a set of student subgroups.

Our first subgroup analysis was by students' grade; we split students into an elementary/middle group and a high school group. We found no meaningful differences in preference by grade group. The only difference in order occurred at Level 3, where older students' top preference was "meeting expectations"; however, it remained the case that these students also preferred "on-track" to the current wording.

Our second subgroup analysis broke down preference by students' race/ethnicity. Due to sample size, we were able to produce these analyses for the Black, Hispanic/Latino, mixed/multiple race, and white subgroups. Again, we found no meaningful differences in preference by students' race/ethnicity; minor differences in Level 3 preferences appeared, but all groups preferred the alternatives to the current wording. Our only potentially substantive finding was that the Hispanic/Latino subgroup did not have a statistically significant preference for "beginning" over "minimal

understanding" at Level 1, but hypothesis testing found that there was no evidence of a preference in the opposite direction, either.

Caregivers

The number of total times that each label was preferred by a caregiver respondent is listed in Table 11. As above, the current wording used on ISRs is italicized, and an asterisk indicates statistical significance. Again, for every hypothesis test, the *p*-value was in fact 0.001 or below.

Table 11. Overall caregiver preference counts

Achievement level	Label	N preferred
Level 1	In need of support	3,505*
	Not yet meeting expectations	2,837*
	Beginning	2,381*
	Minimal understanding	1,774
Level 2	Approaching expectations	3,161*
	Partially meeting expectations	2,584*
	Developing	2,532*
	Partial understanding	2,206
Level 3	Meeting expectations	2,448*
	On-track	1,675*
	Satisfactory understanding	1,111
Level 4	Exceeding expectations	2,690*
	Extensive understanding	1,615
	Advancing	936

*Significantly different from reference language ($p < 0.001$)

Educators

The number of total times that each item was preferred by an educator respondent is listed in Table 12. As above, the current wording used on ISRs is italicized, and an asterisk indicates statistical significance. Again, for every hypothesis test, the *p*-value was in fact 0.001 or below.

Table 12. Overall educator preference counts

Achievement level	Label	N preferred
Level 1	In need of support	934*
	Beginning	915*
	Not yet meeting expectations	883*
	Minimal understanding	491
Level 2	Approaching expectations	1,025*
	Developing	880*
	Partially meeting expectations	743*
	Partial understanding	577
Level 3	Meeting expectations	770*
	On-track	510*
	Satisfactory understanding	333
Level 4	Exceeding expectations	816*
	Extensive understanding	476
	Advancing	308

*Significantly different from reference language ($p < 0.001$)

Qualitative results

Summary of Findings

As mentioned previously, the qualitative analysis is limited to Levels 1 and 2, as they are most in line with the impetus for this data collection effort. For Level 1, "in need of support" was the most popular descriptor in both the educator and caregiver survey data in addition to the focus group data. Students most often mentioned their preference for "beginning" at Level 1. These counts from the qualitative data match what the quantitative results found, which provides support for our use of the qualitative results to unpack stakeholder preferences at this level. For Level 2, "partially meeting expectations" was the most commonly mentioned preferred label among survey groups; however, focus group participants had this as their third favorite choice (only above "partial understanding"). These counts from the qualitative data match what the quantitative results found, which provide support for our use of the qualitative results to unpack the rationales for the indicated stakeholder preferences.

Overall, students had a stronger preference for encouragement relative to clarity. Nevertheless, students mentioned clarity as it related to matching the language used in the labels to their self-perceptions of ability on the test and in school. Caregivers had a slightly larger concern for encouragement than clarity, but they more frequently mentioned clarity in their rationales than the educator and student groups did. Educators preferred encouraging language when they mentioned how the labels would be viewed from a students' perspective, and they preferred clear language when they mentioned how caregivers would interpret assessment results. Subthemes that emerged from the main clarity theme include the following: (1) Direct and Clear Language; (2) Student Progress & Understanding; (3) Unclear Language; (4) Student Needs Help; (5) Stakeholder Action; Expectations; (6) Student Not Understanding; and (7) Other. The subthemes that emerged from the main encouragement theme include: (1) Encouraging and Positive Language; (2) Learning as a Process; (3) Harsh and Negative Language; (4) Student Understanding; (5) Getting Support; (6) Expectations; and (7) Other. All subthemes are listed according to the frequency in which they occurred across all excerpts with the most frequently occurring subthemes listed first. The most prevalent subthemes from the two main themes are further disaggregated by respondent group and discussed below.

Findings by stakeholder group

Students

Among student survey responses, the most common clarity subtheme was "Direct and Clear Language" followed by "Student Progress and Understanding." Students who referenced "Direct and Clear Language" rationalized that the wording they preferred was easier to understand and less vague about what they were able to accomplish. One student wrote, "Approaching expectations, because its more comfortable and its more straight forward when saying that I'm getting there but I'm not quite there yet." This quote shows that the "Direct and Clear Language" subtheme is directly related to students' preference for labels that placed them on a progression of learning. Words and phrases in the "Student Progress and Understanding" subtheme described students' desire for knowing where they were located at that moment in time on a learning trajectory, such as in the

response, “i chose beginning because i feel as if it sets the student up to know were they want to be and were there at…” Their survey responses suggested that they wanted the labels to clearly indicate what they understand and what they have left to learn. Students stated that in order to improve, they wanted to understand clearly that they were in the beginning stages of learning a concept.

Students most frequently mentioned “Encouraging and Positive Language” followed by “Harsh and Negative Language” as subthemes under the encouragement main theme. All stakeholders, but students in particular, often described the tone of one of the indicators to be more encouraging and positive relative to negative connotations in the other choice. One student wrote, “I chose beginning because that is giving the child something to feel good about and minimal understanding just sounds rude and I wouldn’t feel good hearing that.” This excerpt shows that students felt that some indicators used harsh or negative language, which triggered feelings that the student knew very little about the material on the assessment or the subject that was being assessed. Students mentioned that they would feel disheartened if they were to receive a score report with indicators they associated with negative language. Another student wrote, “I choose ‘beginning’ because when you see the words ‘minimal understanding’ it kind of makes you really mad and upset about it.” Across all three quotes reported above, we see that students’ preference for “beginning” as the Level 1 indicator was driven by their preference for language that suggests learning as a process and is encouraging about their potential, especially relative to “minimal understanding,” as they found that language to trigger negative emotions.

Among student excerpts that were coded as “Other,” because they did not directly relate to either encouragement or clarity, students often described how it was important that the labels matched their self-perceptions of how they did on the test or in school. This rationale was associated with multiple Level 1 and Level 2 choices, as students identified multiple ways in which their varying academic experiences were captured in the choices presented in the survey. For example, one student wrote, “I chose the word minimal understanding because I personally did not understand certain matters during the test but there were certain parts on the test that I clearly understood.” This quote exemplifies the feeling many students shared that they wanted the label to reflect their experiences and perceived abilities.

Caregivers

Across levels, the most common clarity subtheme was “Direct and Clear Language,” with “Student Progress and Understanding” the next most frequently coded. These results are analogous to what was reported above for students. Caregivers described labels as giving them “concrete” and “direct” information about their students. As with students, this subtheme was directly related to “Student Progress and Understanding.” Caregivers frequently identified whether the proposed wording better described where a student was performing relative to the original label. Excerpts coded with this clarity subtheme were often related to the preference for labels that referenced “understanding” and “expectations” relative to “beginning” or “developing.” Caregivers often found the latter two choices to be too vague. Additionally, caregivers proposed a paradox by asking if indicators could be both clear and encouraging. They mentioned that the clearest descriptions were not usually the most encouraging, for example, “I’ll just say that in this, in this whole exercise in the survey it struck me that the things that are most positive and encouraging aren’t necessarily the clearest. And so you’re going to lose some clarity by trying to create encouraging messages and that concerns me.” This quote indicates the caregiver’s preference for clarity at the sacrifice of encouragement.

The most prevalent encouragement subtheme among caregivers was "Learning as a Process." Caregivers next most frequently alluded to "Encouraging and Positive Language" in their excerpts coded for the encouragement subtheme. Like students, caregivers noticed the tone of their preferred label was more encouraging and positive relative to the perceived negative connotations they felt in the other choice. Caregivers associated some indicators with a growth mindset and felt they were encouraging because they suggested learning was a process. One caregiver wrote, "I choses approaching expectations because I feel that implies that my child is heading in the right direction and on the right track. Partial understanding sounds like my child could be at a stand still and is not making progress." These respondents felt that phrasing the label to suggest future progress is expected would be interpreted as encouragement.

Caregivers also mentioned the limitations of what a test can tell them about their students: "I do think this is just a test, and that doesn't necessarily accurately identify what students understand I'm not just simply whether or not they can read the question properly and figure out the answer to the test not whether or not they fully understand the concept." We believe that this indicates that many caregivers realize that the state assessments are limited in terms of specific information they can provide about individual students, which indicates a potential need for the state to clarify the purpose and uses of the state assessment with caregivers. This recommendation is further detailed in the final section of this report.

Educators

The most common subtheme coded among the clarity excerpts for educators was "Direct and Clear Language." This matched what was found for students and caregivers. However, educators also frequently mentioned "Unclear Language" as opposed to the student progress subtheme found in the other groups. Educators were particularly concerned that unclear language would be difficult for caregivers to understand. One educator wrote, "Parents do not understand when we sugar coat things. It is important they know their student is not yet where we want them to be." Educators associated unclear language with many of the choices, referencing the fact that either expectations, support, or understanding was not clearly defined within the label. Focus group data revealed that some caregivers had similar feelings.

The most common encouragement subtheme was "Learning as a Process" followed by "Encouraging and Positive Language." The encouragement subtheme findings were very similar to what was found for caregivers. Educators were often concerned that language should be encouraging for students and used words such as "hopeful" and "strengths-based" to characterize labels they preferred. Like caregivers, educators frequently associated encouragement with a growth mindset and frequently mentioned that students should see that they are making progress toward a goal. One educator said, "Again, 'partial understanding' gives a feeling of an end result. 'Approaching expectations' gives a feeling that the journey is still underway: the student can still meet the expectation in the future."

Conclusion and recommendations

The findings from our quantitative and qualitative analyses largely converge. Level 1 is the only place where differing priorities of students, caregivers, and educators have implications for our language recommendations. As noted above, all three groups preferred the label "beginning" to the current phrasing. In light of the findings above, we recommend the following labels for Maryland's achievement levels:

- Level 1: Beginning (in need of support)
- Level 2: Approaching Expectations
- Level 3: Meeting Expectations
- Level 4: Exceeding Expectations

These recommendations are well supported by the findings of our quantitative and qualitative analyses. The qualitative analyses did surface concern among caregivers and educators that the term "beginning" is not sufficiently clear or that it does not sufficiently indicate that a student at Level 1 may benefit from additional support. As such, we recommend including the language "(in need of support)" in parentheses next to the Level 1 label of "beginning."

We also strongly recommend developing additional interpretive supports for caregivers and educators that provide information related to the intended meaning of the achievement levels and any appropriate actions that may be warranted due to the student's score. We suggest including a description of each of the performance levels on the ISRs, as well as adding or enhancing the available score report interpretations guides. This is especially important for those students scoring below proficiency. For example, for students scoring within the lowest achievement level, additional interpretive language for caregivers could be the following: "Your child is scoring in the lowest level measured by the state assessment and is demonstrating *below grade level* knowledge and skills. *Additional and immediate support is needed* to ensure your child is on track for college and career readiness." Supporting language like this can serve as a call to action for parents and educators to advocate for or provide immediate and needed educational interventions for students who are at risk of graduating without the necessary knowledge and skills for meaningfully engaging in their post-secondary plans. While the state assessment is not able to provide detailed information at a granular size that would be useful for informing the exact type of educational intervention needed (e.g., where gaps or misunderstandings may exist), the state assessment can and should be a strong signal to parents and educators about the degree to which the student's current educational program is adequately meeting their needs for reaching the state's grade level expectations. By providing this additional context, we expect that Maryland can simultaneously use the kinds of labels that students find most encouraging, which was the highest priority of this evaluation, and provide the context and urgency that caregivers and educators appreciate.

References

Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology, 38*(2), 113–125. https://doi.org/10.1006/jesp.2001.1491

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78*(1), 246–263. https://doi.org/10.1111/j.1467-8624.2007.00995.x

Broda, M., Yun, J., Schneider, B., Yeager, D. S., Walton, G. M., & Diemer, M. (2018). Reducing inequality in academic success for incoming college students: A randomized trial of growth mindset and belonging interventions. *Journal of Research on Educational Effectiveness, 11*(3), 317–338. https://doi.org/10.1080/19345747.2018.1429037

Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences, 113*(31), 8664–8668. https://doi.org/10.1073/pnas.1608207113

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development.* Psychology Press.

Dweck, C. S. (2008). *Mindset: The new psychology of success.* Random House Digital, Inc.

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures.* National Academies Press.

O'Donnell, F. (2020). What's in a label? Unpacking the meaning of achievement labels from tests. *Doctoral Dissertations, 1856*. https://doi.org/10.7275/15558737

Papay, J. P., Murnane, R. J., & Willett, J. B. (2016). The impact of test score labels on human-capital investment decisions. *Journal of Human Resources, 51*(2), 357–388. https://doi.org/doi: 10.3368/jhr.51.2.0713-5837R

R Development Core Team. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinjosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho., C. M., … Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature, 573*, 364–369.

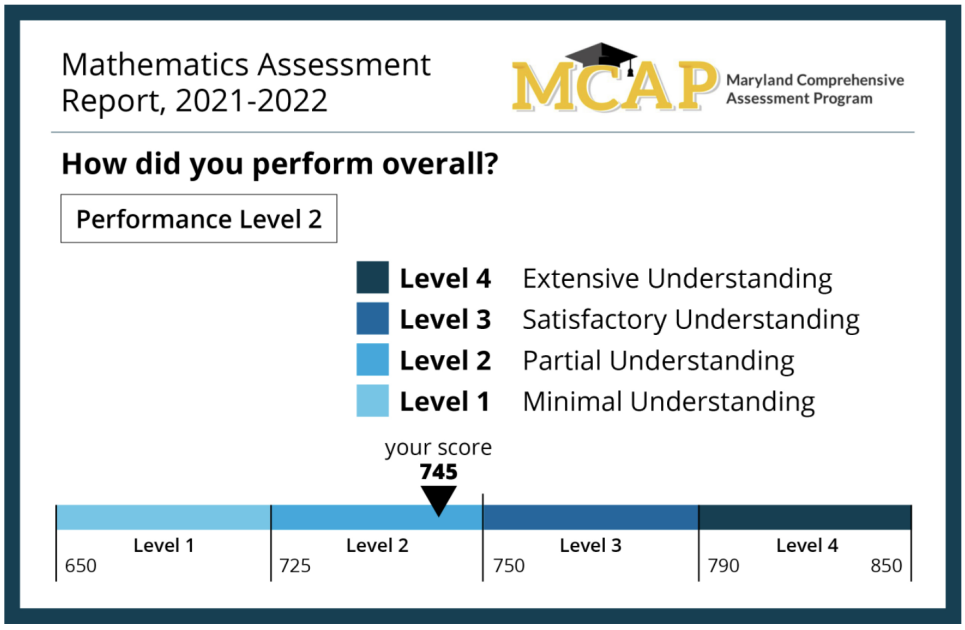
Appendix A: Student survey

Note that items 1–18 appeared in random order on operational surveys, and the choices in each item were also randomized. This counteracts ordering effects. For items 19 and 20, the alternative to the current language presented in the item was random.

Introduction

Thank you for taking the time to answer our survey. This survey asks you to choose between two sets of words that might appear on a score report like the ones below. Please choose the words that you think would make you feel more encouraged if they were on your score report. There are no right or wrong answers to these questions. We are interested in YOUR opinions about how you would feel.

This is an example of part of a score report that a student might receive at the end of the year. Below, you will be asked to compare two sets of words. Imagine that these words were used to describe your performance on the test the way that "partial understanding" was used here.



1. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ On-track
- ◯ Satisfactory understanding

2. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Extensive understanding
- ◯ Exceeding expectations

3. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Approaching expectations
- ◯ Partial understanding

4. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Approaching expectations
- ◯ Developing

5. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Approaching expectations
- ◯ Partially meeting expectations

6. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Minimal understanding
- ◯ In need of support

7. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Partial understanding
- ◯ Partially meeting expectations

8. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Developing
- ◯ Partial understanding

9. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Minimal understanding
- ◯ Beginning

10. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Not yet meeting expectations
- ◯ Minimal understanding

11. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ In need of support
- ◯ Beginning

12. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Meeting expectations
- ◯ On-track

13. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Extensive understanding
- ◯ Advancing

14. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Partially meeting expectations
- ◯ Developing

15. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Advancing
- ◯ Exceeding expectations

16. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Satisfactory understanding
- ◯ Meeting expectations

17. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ Not yet meeting expectations
- ◯ Beginning

18. Which of these descriptions would you find more encouraging to describe your performance on the state test?

- ◯ In need of support
- ◯ Not yet meeting expectations

19. In this survey, you were asked to choose between the terms ‘minimal understanding’ and 'in need of support.' Please explain why you chose the word that you did.

20. In this survey, you were asked to choose between the terms ‘partial understanding’ and 'approaching expectations.' Please explain why you chose the word that you did.

* 21. What grade are you in?

- ○ 5
- ○ 7
- ○ 9

22. What language do you speak the most outside of school? (Optional)

23. What is your race/ethnicity? Please select all that apply. (Optional)

- ☐ White
- ☐ Asian
- ☐ Black or African American
- ☐ American Indian or Alaska Native
- ☐ Hispanic or Latino
- ☐ Native Hawaiian or Other Pacific Islander
- ☐ Other (describe here if you would like)

24. What is your gender? (Optional)

- ○ Male
- ○ Female
- ○ Other (describe here if you would like.)

Appendix B: Caregiver survey

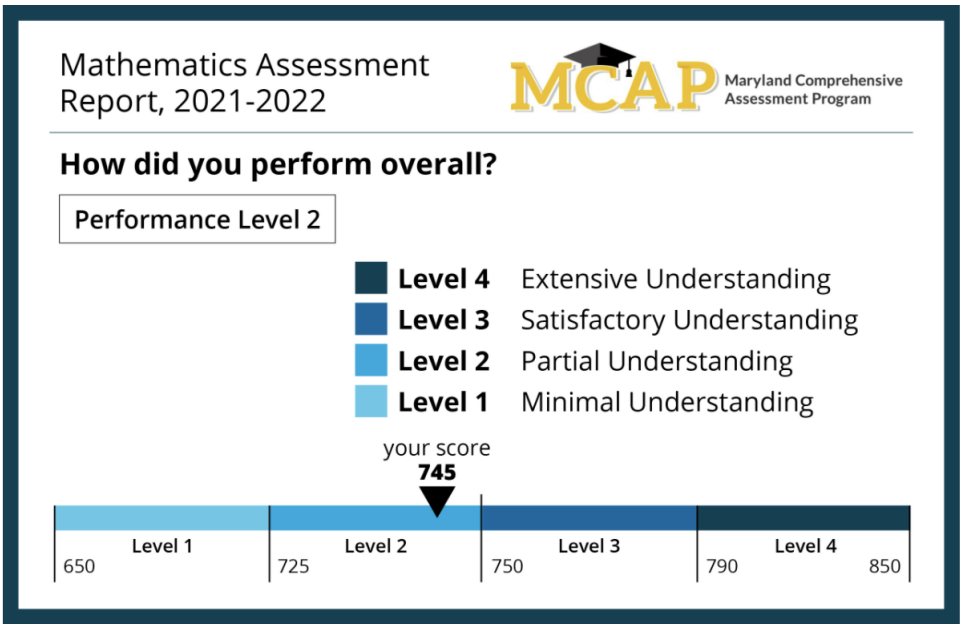
Note that items 1–18 appeared in random order on operational surveys, and the choices in each item were also randomized. This counteracts ordering effects. For 19 and 20, the alternative to the current language presented in the item was random.

Introduction

Thank you for taking the time to answer our survey. This survey asks you to choose between two sets of words that might appear on a score report for your child like the ones below. Please choose the words that you think would make you feel more encouraged about your child's learning if they were on your child's report. We hope to use the results of this survey to improve ISRs for all students and appreciate your participation.

ISR

This is an example of part of a score report that a student might receive at the end of the year. Below, you will be asked to compare two sets of words. Imagine that these words were used to describe your child's performance on the test the way that "partial understanding" was used here.



1. Which of these descriptions would you find more encouraging to describe your child's performance o the state test?

- ◯ Partial understanding
- ◯ Developing

2. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Approaching expectations
- ◯ Partial understanding

3. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Minimal understanding
- ◯ Beginning

4. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Meeting expectations
- ◯ Satisfactory understanding

5. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Satisfactory understanding
- ◯ On-track

6. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Partially meeting expectations
- ◯ Developing

7. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Partially meeting expectations
- ◯ Partial understanding

8. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Not yet meeting expectations
- ◯ In need of support

9. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Minimal understanding
- ◯ Not yet meeting expectations

10. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ On-track
- ◯ Meeting expectations

11. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Minimal understanding
- ◯ In need of support

12. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Approaching expectations
- ◯ Developing

13. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ In need of support
- ◯ Beginning

14. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Partially meeting expectations
- ◯ Approaching expectations

15. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Advancing
- ◯ Exceeding expectations

16. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Advancing
- ◯ Extensive understanding

17. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Beginning
- ◯ Not yet meeting expectations

18. Which of these descriptions would you find more encouraging to describe your child's performance on the state test?

- ◯ Exceeding expectations
- ◯ Extensive understanding

19. In this survey, you were asked to choose between the terms ‘minimal understanding’ and 'beginning.' Please explain why you chose the wording that you did.

20. In this survey, you were asked to choose between the terms ‘partial understanding’ and 'developing.' Please explain why you chose the wording that you did.

* 21. What grade is your child in?

- ○ 5
- ○ 7
- ○ 9

22. What language do you speak the most at home? (Optional)

23. What is your race/ethnicity? Please select all that apply. (Optional)

- ☐ White
- ☐ Asian
- ☐ Black or African American
- ☐ American Indian or Alaska Native
- ☐ Hispanic or Latino
- ☐ Native Hawaiian or Other Pacific Islander
- ☐ Other (describe here if you would like)

24. What is your gender? (Optional)

- ◯ Male
- ◯ Female
- ◯ Other (describe here if you would like.)

25. Would you like to provide more feedback via a small focus group meeting?

- ◯ No
- ◯ Yes

26. If yes, please enter your email address.

27. Please select which date/time option works best for your schedule to participate in a small focus group meeting if interested.

- ◯ 3/14/2022, 12:00-12:45pm
- ◯ 3/14/2022, 5:15-6:00pm
- ◯ 3/14/2022, 7:00-7:45pm
- ◯ 3/16/2022, 12:00-12:45pm
- ◯ 3/16/2022, 5:15-6:00pm
- ◯ 3/16/2022, 7:00-7:45pm

Appendix C: Educator survey

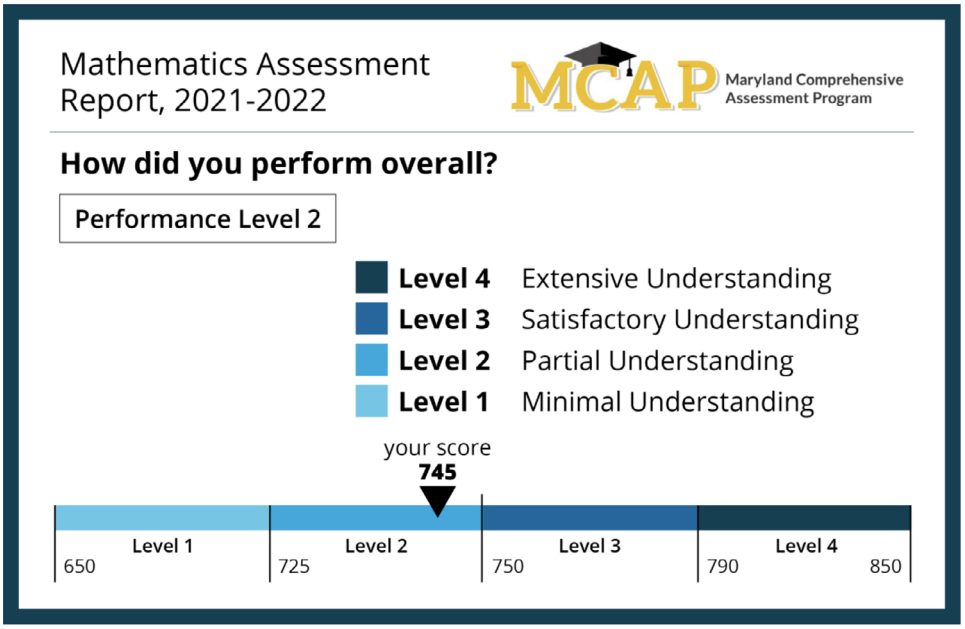
MSDE Assessment Reporting Evaluation: Educator Survey

Introduction

Thank you for taking the time to answer our survey. This survey asks you to choose between two sets of words that might appear on a score report for a student. Please choose the words that you think would make you feel more encouraged about your student's learning. We hope to use the results of this survey to improve Individual Student Reports for all students. We greatly appreciate your participation.

MSDE Assessment Reporting Evaluation: Educator Survey

This is an example of part of a score report that a student might receive at the end of the year. Below, you will be asked to compare two sets of words. Imagine that these words were used to describe your student's performance on the test the way that "partial understanding" was used here.



1. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Partial understanding
- ◯ Developing

2. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Partial understanding
- ◯ Partially meeting expectations

3. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Extensive understanding
- ◯ Advancing

4. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Approaching expectations
- ◯ Partial understanding

5. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Not yet meeting expectations
- ◯ In need of support

6. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Minimal understanding
- ◯ In need of support

7. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Minimal understanding
- ◯ Not yet meeting expectations

8. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Satisfactory understanding
- ◯ Meeting expectations

9. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Developing
- ◯ Approaching expectations

10. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Extensive understanding
- ◯ Exceeding expectations

11. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ On-track
- ◯ Satisfactory understanding

12. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Exceeding expectations
- ◯ Advancing

13. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Beginning
- ◯ Minimal understanding

14. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Meeting expectations
- ◯ On-track

15. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Developing
- ◯ Partially meeting expectations

16. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Beginning
- ◯ In need of support

17. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Beginning
- ◯ Not yet meeting expectations

18. Which of these descriptions would you find more encouraging to describe a student's performance on the state test?

- ◯ Partially meeting expectations
- ◯ Approaching expectations

19. In this survey, you were asked to choose between the terms ‘minimal understanding’ and 'beginning.' Please explain why you chose the wording that you did.

20. In this survey, you were asked to choose between the terms ‘partial understanding’ and 'partially meeting expectations.' Please explain why you chose the wording that you did.

MSDE Assessment Reporting Evaluation: Educator Survey

Background information

21. Which grade(s) do you teach?

- [ ] K
- [ ] 1
- [ ] 2
- [ ] 3
- [ ] 4
- [ ] 5
- [ ] 6
- [ ] 7
- [ ] 8
- [ ] 9
- [ ] 10
- [ ] 11
- [ ] 12
- [ ] not listed

22. What language do you speak the most at home? (Optional)

23. What is your race/ethnicity? Please select all that apply. (Optional)

- [ ] American Indian or Alaska Native
- [ ] Asian
- [ ] Black or African American
- [ ] Native Hawaiian or Other Pacific Islander
- [ ] Hispanic or Latino
- [ ] White
- [ ] Not listed (describe here if you would like)

24. What is your gender? (Optional)

- ◯ Female
- ◯ Male
- ◯ Not listed (describe here if you would like.)