Determining Student Impact Implementation Brief

Investigating Fairness

Question: *The focus question for this brief.*

How do I ensure DDMs are fair assessments of student growth and educator impact?

District Readiness: *Work to be completed before answering the focus question of this brief.*

- ☑ District has engaged educators in the DDM identification/development process, including educators of students with disabilities, ELLs, and other special populations.
- ☑ District is ready to pilot or administer assessments that are aligned to content and provide useful information to educators about student learning, growth, or achievement.

Next Steps: *Suggested next steps for districts after reading this brief.*

- ☐ Develop long-term plan to investigate issues of fairness
- ☐ Develop long-term plan to modify or replace DDMs as needed.

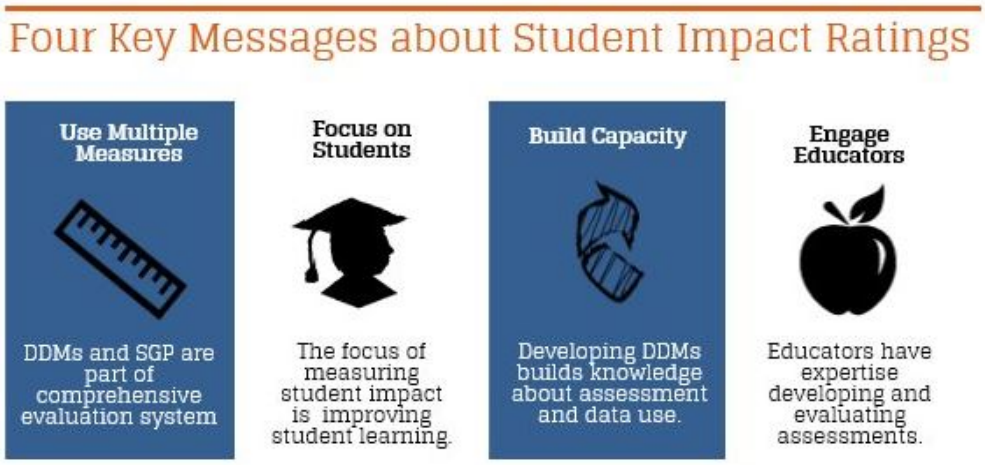
DDM Development

This Implementation Brief is relevant to the highlighted steps:

- Selecting
- Administering
- Scoring
- Analyzing
- **Adjusting**

As districts pilot and implement DDMs they often face questions about fairness when using DDMs as a part of educator evaluation. This brief prepares districts to investigate and address issues of fairness in DDMs. Districts should make a good faith effort to investigate and address potential fairness issues moving forward. However, concerns about fairness should not paralyze districts from continuing to identify and implement measures of student growth. This brief introduces, several sophisticated concepts and approaches, many that are designed to investigate issues after DDMs have been implemented. These concepts and approaches are introduced here to support future planning, but may not be timely or necessary for all districts. For more information about fairness watch Webinar 6 of ESE's DDM and Assessment Literacy Webinar Series.[i]

These Implementation Briefs are designed to provide targeted guidance focused on timely questions around the implementation of District Determined Measures. These briefs highlight important questions, resources, and approaches for districts to consider, and are not exhaustive resources on the subject. Please continue to share your own examples, suggestions, and questions with us at EducatorEvaluation@doe.mass.edu.



Investigating Fairness

Understanding Error:

No assessment perfectly measures what a student knows. In some situations, this fact should raise concern about the fairness of an assessment, in other situations it should not. Error is the technical term used to refer to the difference between a student's score and his/her true ability. There are two types of error: **random error** and **systematic error**. Random error does not disadvantage any specific group of students or educators. Systematic error does.

Random Error: All assessments have random error. For example, a student may make a couple of lucky guesses and earn a higher score than would be expected, or they may underperform due to a poor night of sleep. The amount of random error in an assessment is what determines an assessment's **reliability**. An assessment with high reliability has low random error, while an assessment with low reliability has high random error. While error cannot be completely eliminated, it can be decreased. Increasing the number of assessments, calibrating raters, or increasing the number of items on a single assessment decreases random error. Since an educator's Student Impact Rating is based on many different students completing multiple assessments over multiple years, random error will not have a strong influence on a teacher's rating (see call out box for further explanation).

Student Impact and Random Error

To understand why increasing the number of assessments or items on a single assessment decreases error, consider rolling a six sided die. If you roll a die once, you have an equal chance of rolling each number, 1-6. However if you were to keep rolling, the average value of your rolls would begin to converge to the die's "true average value" of 3.5. We could say, then, that a single die roll has a high amount of error for estimating a die's "true average value." However, if we take the average of many different die rolls, sometimes the roll would be high and sometimes it would be low, but the average will be close to the "true average value" of 3.5. Similarly, increasing the number of assessments or items on a single assessment will decrease the likelihood that random error will cause a student's true ability to be misrepresented. If we extrapolate this principle to educator impact, the fact that a Student Impact Rating is based on multiple assessments administered to many students over at least two years mitigates the potential impact of random error.

Systematic Error: Systematic error is a special type of error that is not random. This means that one or more groups of students are more likely to do worse (or better) on assessment than one might expect given their knowledge and skills. This is the technical definition for **bias** in an assessment. For example, students may not be able to demonstrate high growth because of high initial (e.g., pre-test) scores. In this case, the assessment would underestimate the growth of students who had previously performed well. As a result, an educator who worked with these students would not be able to demonstrate high impact, even if students had made tremendous gains in understanding the material. Systematic error may be small in magnitude, but unlike random error, systematic error is not decreased by increasing the number of assessments or the number of items on a single assessment. As a result, systematic error can play a greater influence than random error on a Student Impact Rating. This is one reason why districts will pay particular attention to mitigating the effects of systematic error when investigating issues of fairness.

Investigating Fairness

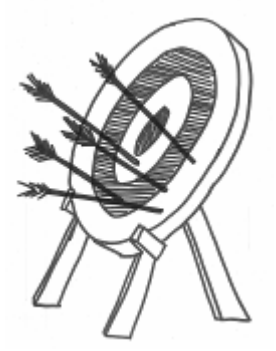
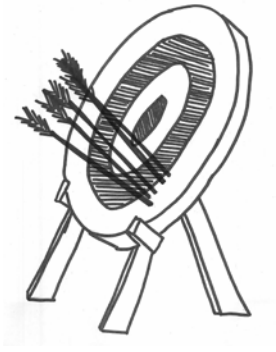
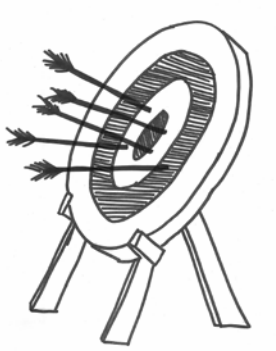
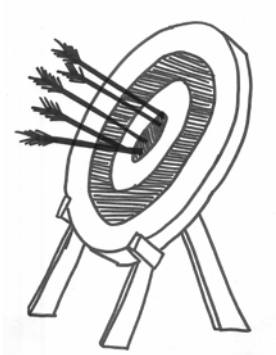
		Random Error: High	Random Error: Low
Systematic Error (Bias)	High		
	Low		

Figure 1.

In the four pictures to the left we see examples of targets that represent high and low systematic and random error. Using the bullseye example, think about systematic error as the assessment's level of inaccuracy and random error as its level of imprecision.

The **top left** picture is an assessment that is **both inaccurate and imprecise** – not a good DDM. The **top right** picture is also **problematic** because while all arrows have landed in a simalr place, they are far off the mark. High systematic error in an assessment means that what is being measured (i.e., a student's ability to read a complex word problem in English) is not what is intended to be measured (e.g., a student's ability to demonstrate understanding of fractions).

Conversely, the **bottom right** picture is **both accurate and precise** – an ideal DDM. And while the **bottom left** is an assessment that can be improved, it does not unfairly disadvantage a specific population (i.e., every archer has a chance at the bullseye).

What this means for DDMs: When implementing DDMs it can be overwhelming to consider all the different sources of error. However, districts should be reassured that not all sources of error are equally important. As long as the potential sources of error are not concentrated for one educator or group of students, then it is unlikely to cause an issue of fairness. For example, a confusing question may introduce considerable random error, but if there is no reason to believe that one group of students is more likely to make a mistake, this error does not translate to an issue of fairness. However, systematic differences, such as misreading a question, may occur more often to one group of students. Addressing this type of error is where districts should focus their attention.

Investigating Fairness:

Professional Judgment: Districts should engage educators in the process of investigating the fairness of DDMs. One of the benefits of implementing comparable assessments is the ability to then engage role-alike educators in meaningful conversations about student growth. Such conversations are made more robust by engaging educators in the process of investigating issues of fairness. A starting place for this work is to provide opportunities for educators piloting or implementing the same DDMs to collectively review the assessments and use their professional judgment to flag questions or prompts that are likely to advantage or disadvantage

certain groups of students. This type of exercise is an appropriate entry point for educators to begin to think about fairness. The following section describes some quantitative methods for investigating issues of fairness that some districts may be poised to consider.

Checking Correlation: Districts that are interested in investigating bias after a DDM has been administered can look at the results to see if one type of student is less likely to demonstrate growth than other students. While different groups of students may demonstrate different average levels of achievement, all groups of students should have an equal chance of demonstrating growth on a DDM. If the results show that one type of student has a reduced (or increased) chance of demonstrating growth then an issue of bias has been uncovered at the student level. One common issue that might be revealed is that students who scored low on a baseline measure demonstrated less growth on the assessment. This is an example of systematic error, because one group of students, in this case lower performing students, was more likely to demonstrate low growth. Districts can look for this issue by checking to see if there is a relationship between students' starting, or baseline, scores and their growth scores.

Correlation is a technical way to check the degree to which two different numbers are related that can be used to determine if a DDM has systematic error. You can learn more about correlation here.[ii] One way to use correlation to investigate systematic error is to ensure that there is **no relationship** between students' baseline performance (i.e., pre-test scores) and their growth scores. It is important to remember that a relationship between a student's pre-test and post-test score is not a problem since we would expect that students that scored higher than most students on the pre-test would score higher than most students on the post-test. However, if the DDM is a fair assessment of student growth, students with low pre-test scores will demonstrate growth at roughly the same levels as students with moderate and high pre-test scores. Scatter plots are a visual approach for exploring the correlation between pre-test scores and gain scores. Figures 2 and 3 illustrate two examples of how scatter plots can help one check for systematic error.

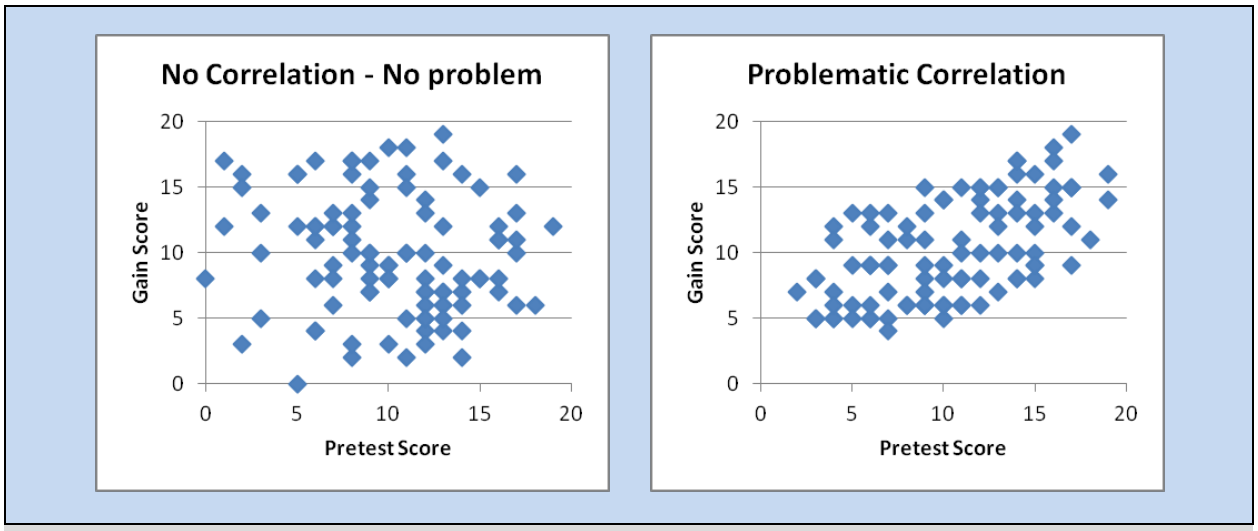


Figure 2. In the example on the left, students who scored below a 5 on the pre-test mostly had gain scores between 5 and 17. Students who scored above 15 had a similar distribution, providing evidence that there is no systematic error in this assessment. The graph on the right is problematic. Students who scored low on the pre-test had gain scores between 3 and 13, while students with high pre-test scores had gain scores between 10 and 20. As a result, not all students had an equal chance of demonstrating growth on this DDM

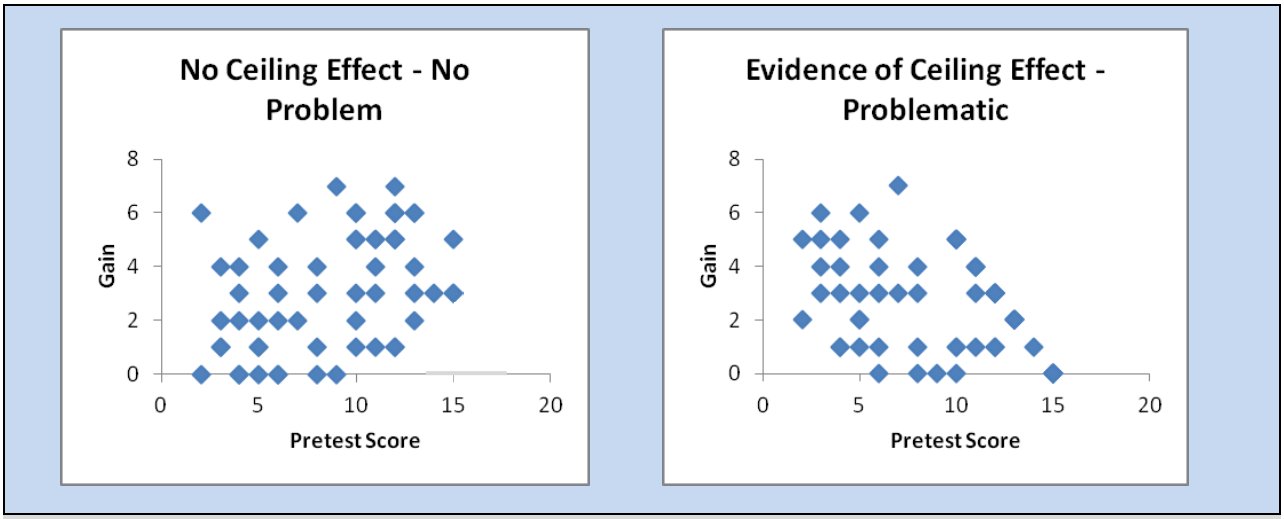


Figure 3. The graph on the left shows another example of an assessment that yielded no relationship between students' pretest and gain scores. In the graph to the right, there is evidence of systematic error due to a ceiling effect. That is, some students earned the maximum score (the ceiling) on the post-test even when they may have been able to demonstrate greater learning. Because the ceiling was not a factor for students with low and moderate pre-test scores, these students were able to demonstrate more growth. The ceiling effect harmed students who scored high on the pre-test because the amount of growth they could demonstrate was capped by the number of available points in the assessment scoring scale. For example, using a 15 point scale, a student who scored a 14 on the pre-test will only be able to demonstrate one point of growth. The scatter plot on the right clearly shows the ceiling in the form of the diagonal line of data points showing where students were limited in the amount of growth they could show based on high pre-test scores.

Checking for Variability: A lack of comparability between DDMs is another potential issue of fairness that districts may encounter. One way to check for a potential issue is to see if different DDMs are yielding similar variability in student results. Variability, in the context of DDMs, is the degree to which a measure identifies similar numbers of students as demonstrating high, moderate, and low growth. Districts may want to use bar graphs to compare the number of students identified as demonstrating high, moderate, or low, such as in the figure 4 below. Looking at this distribution across multiple DDMs is one way to draw conclusions about whether DDMs are comparably rigorous across content areas. For example, if one assessment is more likely to identify students as having low growth, simply because it is more rigorous than others, and not because students are actually demonstrating less growth, an educator using this DDM would be disadvantaged relative to his/her peers.

Variability is also important because it ensures that DDMs provide meaningful information about students. For example, let's say a district works with educators to set parameters for which bands of student scores translate to high, moderate, and low growth for a particular DDM. The DDM is then administered and scored and the district finds that all student scores fell into the "moderate" band. In this case, the educators using the DDM have learned very little about their students because the assessment failed to differentiate student growth. While there is no requirement to set a fixed number of students identified as having high, moderate, and low growth, DDMs that consistently identify all or most students in the same category actually provide little actionable information about student growth.

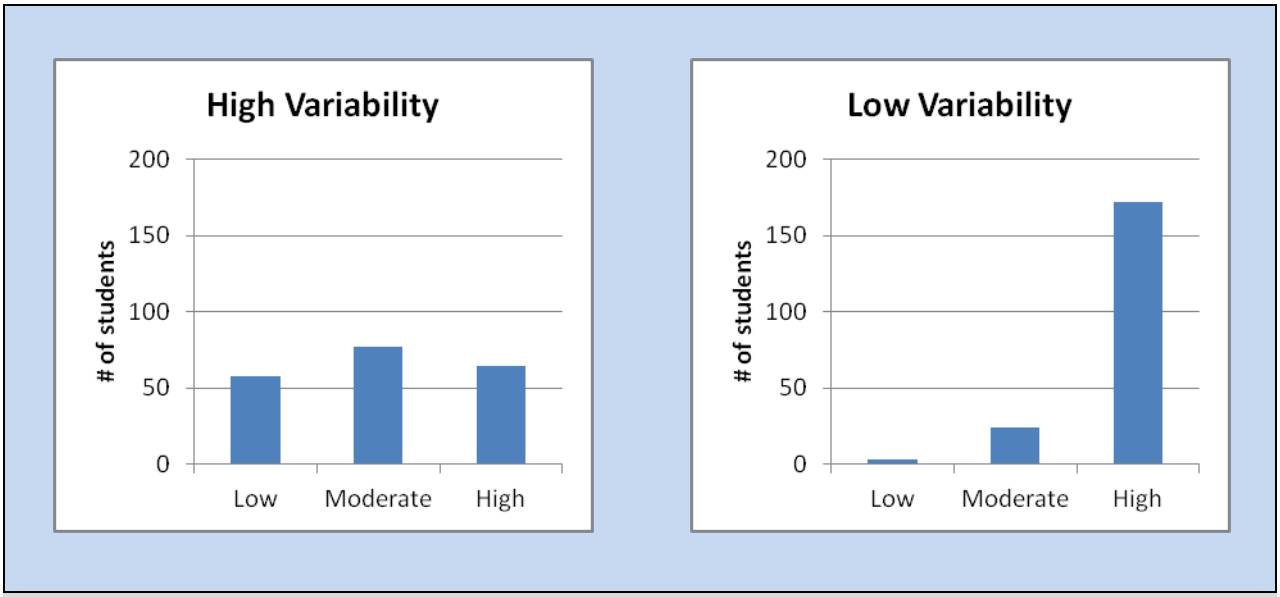


Figure 4. The figure on the left provides an example of significant variability between student scores. Enough students are identified in each group that this DDM will provide a chance to foster positive discussions around how to help these different groups of students. In the example on the right, this DDM provides little distinction between students demonstrating growth.

To be clear, setting rigid percentages of students that must be identified as demonstrating high, moderate, and low growth is not recommended. Districts should always use their best professional judgment. For example, if a DDM had variability in previous years, or if a measure is based on national norms (e.g. growth norms from MAP), a district might have confidence that the lack of variability in a single year was not due to the measure itself, but was an accurate reflection of student growth within a particular cohort of students. However, for many DDMs, variability will be an important piece of information to consider during the parameter setting process. For more information see the Implementation Brief on Scoring and Parameter Setting.[iii]

Addressing Issues of Fairness

The previous section described strategies for investigating whether a DDM is fair. This section provides districts with some techniques for mitigating the consequences of an unfair DDM. Some of these techniques involve data analyses that some, but not all districts may be equipped to pursue. The ideas discussed here are not required components of a district's implementation efforts, but rather food for thought when deciding how to address issues of fairness that may arise.

Districts should plan for the continuous improvement of DDMs over time. Part of this continuous improvement work will be focused on making improvements to DDMs that initial data suggest may produce significant systematic error. Sometimes, educators can make small modifications to the assessment that will go a long way toward reducing systematic error. For example, say a district uses one of the strategies described above to investigate fairness and finds that an assessment does not provide lower performing students an opportunity to demonstrate growth at the same rates as their higher performing peers. One can imagine an extreme scenario

where a student scores a zero on a pre-test and a zero on the post-test. The student may have learned a great deal during the instructional period, but the assessment is too blunt to pick up on this growth. The educators working with the DDM might try adding additional easy questions to both the pre- and post-tests, which will provide lower performing students opportunities to answer more questions correctly. As a result these students will be more likely to demonstrate growth at similar rates as other students.

Systematic error may also be reduced by removing problematic questions. For example, if a question contains an unnecessary word that disadvantages one cultural group, this question should be edited. Analyzing student responses on a question-by-question level may help reveal problematic questions that can be modified or discarded.

Adding questions of varying difficulty and removing problematic items are two ways to reduce systematic error that involve making changes to the actual assessment. However, sometimes it will not be possible to address systematic error in these ways. Below are three additional approaches for addressing issues of fairness that involve interpreting student responses rather than making modifications to the assessment.

Example: How Grouping Addresses Bias

Teacher A

Pre	Post	Gain
3	4	1
3	4	1
3	4	1
3	4	1
8	14	6

average gain is 2

Teacher B

Pre	Post	Gain
3	4	1
8	14	6
8	14	6
8	14	6
8	14	6

average gain is 5

Average Gain By Group

	Teacher	Average Growth
Low-Performing Students	A	1
	B	1
High Performing Students	A	6
	B	6

no differences in average gain

Figure 4. This demonstration shows how grouping can address bias at the student level. Teacher A and B administer the same assessment to their classes. All of the students who scored a 3 on the pre-test scored a 4 on the post-test. Similarly, all of the students who scored an 8 on the pre-test scored a 14 on the post-test. Looking only at the average gain score by class, one might conclude that Teacher B had a higher impact on his/her students because his/her students' average gain was 5, whereas Teacher A's students' average gain was 2. By using grouping, a district can see that, in fact, Teacher's A and B had an equal impact on their students. It just so happened that Teacher A's class was comprised of a greater number of lower performing students.

Comparable Assessments: Districts are not required to implement identical DDMs within a grade/subject or course. The regulations require that DDMs be *comparable*. As a general rule, using identical DDMs for educators in the same or similar roles supports fairness and comparability. However, there may be situations when identical DDMs are not the most

603 CMR 35.09(2)a.[1] DDMs must be *"comparable across schools, grades, and subject matter district-wide."*

appropriate option. If a given assessment produces systematic error at the student level, meaning students do not have the potential to demonstrate the same amount of growth, choosing an appropriate alternative, but comparable instrument may provide the best opportunity to address this systematic error. For example, consider a math test that is found to be unfair for ELL students because of the reading skills necessary to respond to the questions. While reading skills are important, the math test would not be a fair assessment of student growth *or* educator impact if the educator was not responsible for teaching the reading skills required to comprehend the math test. In fact, since ELL students have to master two different skills to answer questions correctly, the test may be twice as difficult for these students. A version of the math test translated into the students' native languages, may result in a comparable, albeit not identical, assessment of student growth, and therefore, educator impact. Another example of a scenario where a district might decide to use comparable, but not identical assessments is when two educators teach the same course, but at different levels. A district may have a teacher who teaches a basic Algebra I class and another who teaches an advanced Algebra I class. It may not be possible to select a single DDM that provides an equal opportunity for all students in these two classes to demonstrate growth.

Grouping: Grouping is another strategy for interpreting student scores that can address potential systematic error at the educator level. When we group, we compare a student's growth score to the growth scores of similar students. If lower ability students receive lower growth scores, solely calculating a median student growth score may not be fair to an educator who works primarily with low performing students. However, a fairer comparison of educator impact might be possible if we group like-performing students together and look at the median growth score for each group. With this strategy it does not matter how many students of each type a teacher is responsible for. It is important to realize that the assumption behind using a grouping strategy is that the measure of growth produces systematic error for one group of students. Grouping reflects an effort to mitigate the effects of this systematic error. Grouping is not used to set different expectations for different types of students. Figure 4 presents an example of how grouping can reveal aspects of an educator's impact that would be lost by only calculating average gain scores. Webinar 6 of ESE's DDM and Assessment Literacy Webinar Series provides further explanation of how to use grouping effectively.

If a district chooses to use grouping, they should keep the number of groups small enough so comparisons are feasible, and reflect enough students to make reasonable comparisons. Grouping is a reasonable choice when working with commercial assessments that are difficult to modify, or to address systematic error revealed after an assessment has been administered.

Standardization: Ensuring that different assessments are comparable can be challenging. One way to make comparisons easier between different assessments is a process called standardization. Standardization involves placing numbers on the same scale, based on their relative position in a normal distribution. Standardizing assumes that most of the numbers in a test are in the middle, with fewer cases of very low and very high numbers. The closer a standardized number is to zero, the closer it is to the average. For example, if a student had a

standardized score of one on an assessment, then we know that this student performed one standard deviation above the mean, or roughly better than about two-thirds of other students, even without knowing anything about the assessment. Webinar 6 provides further information about standardization and how one district used it to support DDMs.

Frequently Asked Questions

Do all students in like classes need to be assessed by the same DDMs? No. As with all DDMs, districts should be guided by the key questions outlined in Technical Guide B[iv]: A DDM must assess what is most important for students to learn and be able to do and what the educators intend to teach. A DDM must also provide valuable information to educators about their students and to the districts about their educators. If a DDM doesn't meet these two goals for all like classes, districts might consider using different DDMs.

How do I balance fairness with feasibility? Districts need to take fairness and systematic error in DDMs seriously. However, it may not be possible for districts to create DDMs completely free from systematic error right away. Investigating and addressing systematic error should be part of districts' plan for the continuous improvement of DDMs. Information about systematic error may also be considered by evaluators in the application of professional judgment in determining an educator's Student Impact Rating.

[i] http://www.doe.mass.edu/edeval/ddm/webinar.html
[ii] http://www.mathsisfun.com/data/correlation.html
[iii] http://www.doe.mass.edu/edeval/ddm/Scoring&PSetting.pdf
[iv] http://www.doe.mass.edu/edeval/ddm/TechnicalGuideB.pdf