Tips for Using Data Responsibly*

Guiding Questions for Responsible Data Use

- ***Validity:*** What are my intended inferences and uses of this assessment? What evidence supports these interpretations? Does the available evidence convince me that my inferences are reasonable and well-supported?
- ***Reliability:*** If I replicated this assessment, how variable would the scores be? Do I have enough data points for the scores to be generalizable?
- ***Sampling:*** Am I focusing on the "easy to measure" or "high stakes" skills, or am I covering the full domain of desired outcomes?
- ***Triangulation:*** Do other data sources support my inferences and uses of this assessment?

Types of evidence you can gather to evaluate the validity of your inferences:

1. Reliability data
2. Analysis of the content of the test
3. Statistical analysis of performance on the test
4. Relationships between scores and other variables
5. Responses of students taking the test
6. Sampling and measurement

Evidence of Validity #1: Reliability of Scores

- **Provide multiple equivalent assessment opportunities** to evaluate student skills. This will help ensure your conclusions are generalizable. The more raters/ items/ days we can average over, the more stable scores will be.
- **Regularly utilize procedures and practices that foster consistency of measurement**, such as rubric calibration and observation norming. Calibrate grading with colleagues and students. Have multiple raters evaluate responses using a common rubric and try to reach consensus. Norm scoring on constructed response items by comparing against exemplars and colleagues.
- **Carefully construct good test questions.** Phrase each question clearly so that students know exactly what you want. Try to write items that discourage guessing and that discriminate among top-performing and low-performing students and are of an appropriate difficulty level. Reliability is higher when scores are spread out over the entire scale, showing real differences among students.

Evidence of Validity #2: Content of the Test

- **Review the content of tests** to determine how adequately the test measures the intended skills. Consult the relevant content standards, curriculum maps, and pacing guides to gather evidence of alignment. Unpack each assessment item to clarify what skills are needed to answer successfully.
- **Analyze assessment results with test-in-hand** in order to compare results with the actual items. This will enable you to increase the specificity of your evidence and conclusions.
- **Review standards to check for gaps or redundancies in assessment.** Identify under- or over-representation of concepts, untested skills, and limited variation in how tests assess various skills.

*This handout was developed by Mary Dillman, Director of Data Inquiry for Boston Public Schools. It draws on insights contained in: Koretz, Daniel. (2008). *Measuring Up: What Educational Testing Really Tells Us*. ("Validity", pp. 215-234). Cambridge: Harvard University Press.

Evidence of Validity #3: Responses of Students Taking the Test

- **Gather evidence about what students are actually thinking** as they complete a test or task. You can do this by taking your own test and being cognizant of your thinking process – What skills and knowledge were you utilizing? Educators can also observe students to make inferences – pay attention to pace, use of materials, self-revision, non-verbal expression. Or teachers can interview students about what they were doing and thinking as they answered specific questions.

Evidence of Validity #4: Performance on the Test

- **Make critical comparisons to contextualize performance**. For instance, compare a student's progress to the average for the class, school, district, state, nation, etc.
- **Know what performance differences are educationally meaningful.** On state assessments, there is often guidance on how to interpret scores, including what differences are due to measurement error or random chance.
- **Consider each student's or subgroup's distance to typical performance** rather than simply analyzing absolute performance. This will help you understand whether a performance difference on a given item is due to below-average achievement, or variability in item difficulty.

Evidence of Validity #5: Relationships between Scores and Other Variables

- **Provide multiple, varied opportunities to evaluate student learning** in order to disentangle student performance on the assessment format from student performance on the skill/concept.
- **Examine how performance on one assessment correlates** with performance on other assessments, including tests that are intended to assess similar domains of knowledge and those intended to assess different domains of knowledge.
- **Vary the "stakes" when assessing the same skill/concept** to account for score inflation and the effect the testing environment/ conditions may have on student performance.

Evidence of Validity #6: Sampling & Measurement:

- **When analyzing a sample of student work, make sure the sample is representative** and reflects the full spectrum of the intended population. For instance, if you want to make inferences about an entire class, select a mix of high-, medium-, and low-quality work artifacts. Understand the performance distribution as well (e.g., proportion of work considered "medium" versus "high").
- **Assess all students in a group** if you want to make inferences about that group's learning.
- **Triangulate performance across multiple assessments to see the full picture.** One of the ways to deal with measurement errors is to use multiple measures of the same concept. Triangulating across multiple measures will provide a more accurate sense of what's going on. Educators should make inferences about student learning based on what the preponderance of evidence indicates.
- **Understand the weight of skills/concepts within each assessment** to guard against under-representation or over-representation. For instance, if you are attempting to assess multiple sub-skills to make inferences about student mastery of a broad concept, make sure the amount of the test dedicated to each sub-skill approximates its proportion of the overall domain.
- **Determine whether a performance task or project requires skills and knowledge outside the domain of skills you are assessing.** (For instance, a math portfolio might actually assess students' writing skills, and successful completion of a science experiment may require fine motor skills.) Complement these performance tasks with other types of assessment.
- **Consider sample size when studying assessment reports.** For test populations smaller than 30 students, differences could be due to random chance.