Technical Guide A, Appendix B. Resources for DDMs

References Cited in Technical Guide A:

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (8th Edition). Upper Saddle River, New Jersey: Prentice-Hall Inc.

Massachusetts Department of Elementary and Secondary Education. (September 2008). *Ensuring technical quality: Policies and procedures guiding the development of the MCAS tests.* Retrieved May 10, 2013 from: http://www.doe.mass.edu/mcas/tech/technical_quality.pdf

Massachusetts Department of Elementary and Secondary Education (n.d.). Massachusetts Comprehensive Assessment System: MCAS Alternative Assessment. Retrieved May 10, 2013 from: http://www.doe.mass.edu/mcas/alt/

National Evaluation Systems, Inc. (1991). Bias issues in test development. Amherst, MA: National Evaluation Systems, Inc.

National Research Council. (1993). *Measuring up: Prototypes for mathematics assessment. Washington, DC: National Academy Press*. Available from: http://www.nap.edu/catalog.php?record_id=2071

Standards for educational and psychological testing. (1999). Washington, DC: American Educational Research Association.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African-Americans. Journal of Personality and Social Psychology, 69, 797-811.

Website Resources for Selecting, Developing, and Evaluating DDMs:

The links in this section are intended for use as resources. Districts may review these sources to generate ideas for DDMs.

Resources for Measures of Grades Pre- K – 2
Early childhood reading assessments in Florida: http://www.fcrr.org/assessment/pdf/microsoft-powerpoint-faculty_presentation_052004.pdf
K-1 literacy and numeracy assessments in Minneapolis:[1] http://rea.mpls.k12.mn.us/uploads/kind_summary_fall_2010_with_2009_bench_2.pdf http://rea.mpls.k12.mn.us/uploads/newer_goa_with_2010_benchmarks_3.pdf
Fountas and Pinnell Benchmark Assessment System, individually administered, one-on-one assessment that matches students' instructional and independent reading abilities to the company's "text level" indicator. http://www.heinemann.com/fountasandpinnell/BAS2_Overview.aspx
Jerry Johns Basic Reading Inventory, Grades K-Adult, http://www.harperhighschool.org/ourpages/auto/2009/2/17/61155853/Prosody.pdf

[1] Minneapolis has also devised a way to create a composite growth score of early literacy skills, complete with cut scores for decision-making related to instructional planning and response to intervention.

Resources for Measures of Grades 3 – 8
Measurements of Student Progress, Grades 3 – 8, Reading, Math, Writing and Science, State of Washington: http://www.k12.wa.us/assessment/StateTesting/MSP.aspx.
New York State Passage Selection Resources for Grade 3-8 Assessments, Grades 3 – 8, CCSS for ELA & Literacy. As such, the tests will approach reading, writing, and language: http://engageny.org/resource/new-york-state-passage-selection-resources-for-grade-3-8-assessments

Resources for Measures of Grades 9 – 12
A list of assessment resources tailored toward geosciences that have some applicability to more content areas: http://serc.carleton.edu/NAGTWorkshops/assess/types.html
A compilation of Common Core assessment resources from the Algebra 1 Teacher's blog: http://blog.algebra1teachers.com/2012/07/common-core-assessment-examples-for.html

Portfolio and Performance Assessment Examples
Tennessee Fine Arts Growth Measures System: http://team-tn.org/assets/educator-resources/TN_ARTS_SCORING_GUIDE_2.5.pdf
Measuring Up: Prototypes for Mathematics Assessment (1993), available from: http://www.nap.edu/catalog.php?record_id=2071
Tutorial for developing a summative performance assessment: http://www.csus.edu/indiv/j/jelinekd/EDTE%20116/116%2009-10/Sessions%20Spr10/Session%2011%20Perf%20Assmt(14apr10).pdf

Additional Guidance to Perform Item Analyses

Item Difficulty

Item difficulty can be directly estimated using students' scores on the assessment. On selection items or other 1-point items, item difficulty can be estimated simply by computing the average score on the item (Excel command: average).

A simple four-item test is used to demonstrate this procedure. The "Difficulty" row is simply the average number of score points earned by the examinees.

Item Difficulty, Four-Item Test				
Examinees	Item 1	Item 2	Item 3	Item 4
1	1	0	3	4
2	1	0	1	4
3	0	0	0	0
4	1	1	4	4
5	1	0	3	4
6	1	1	4	4
7	1	1	3	2
8	1	0	4	4
9	1	0	4	4
10	1	0	1	1
11	1	1	3	4
12	1	0	4	4
Difficulty	0.92	0.33	2.83	3.25
Max. Points	1	1	4	4

Initially, we can see that there are two 1-point items and two 4-point items. All examinees except #3 got Item #1 correct, so this is an easy item with an item difficulty of .92 on this test with this group of examinees (the item difficulty statistics is "sample dependent" and likely to change with a different group of students). Item #2 is much more difficult with an item difficulty statistic of .33.

We can further determine that Item #4 is an easier item (for this group of students) than Item #3 with an average difficulty of 3.25 vs. 2.83; however, we cannot directly compare the item difficulty statistics for the 1-point items to the 4-point items. To compute item difficulty statistics that are comparable, regardless of the number of item score points, we divide the Difficulty statistic by the maximum number of item score points which results in a Difficulty2 statistic that ranges from 0 to 1, or from hard (items) to easy (items).

Item Difficulty, Four-Item Test				
Examinees	Item 1	Item 2	Item 3	Item 4
1	1	0	3	4
2	1	0	1	4
3	0	0	0	0
4	1	1	4	4
5	1	0	3	4
6	1	1	4	4
7	1	1	3	2
8	1	0	4	4
9	1	0	4	4
10	1	0	1	1
11	1	1	3	4
12	1	0	4	4
Difficulty	0.92	0.33	2.83	3.25
Max. Points	1	1	4	4
Difficulty2	0.92	0.33	0.71	0.81

Now we see that Item #1 remains the easiest item with an item Difficulty2 of .92. Item #4 is the next easiest item with a Difficulty2 of .81. Item #2 remains the most difficult item of the group with a Difficulty2 of .33.

Another way to describe difficulty is to say that 92% of the examinees got Item #1 correct; or for the 4-point items, that examinees earned 81% of the score points.

Item Discrimination

Item discrimination describes how well an assessment item discriminates between high- and low-scoring examinees. Positive item discrimination indicates that higher-scoring examinees perform better on the item than lower-scoring examinees. It is important to examine the discrimination of items on common assessments (assessments administered across all students in a district) to check on item quality on the test. Negative item discrimination is undesirable and indicates that lower-scoring examinees performed better on the item than those with higher scores – this is a flag for an item that has a flaw, such as the item:

- has the wrong correct answer indicated
- has two correct answers
- is unclear or confusing to examinees who know the material
- represents different content than the rest of the content on the assessment.

Items with no (“0”) discrimination do not discriminate between high- and low-scoring examinees. These, generally, are items that most or all examinees get either correct or incorrect. In some cases, however, items have no discrimination but there is a mix of students who got the item correct/incorrect. In that case, the item is performing poorly and may be removed from the assessment.

Items with negative discrimination and items with no (“0”) discrimination

We next describe two methods for identifying item discrimination.

The first method begins by dividing the examinees into two halves, based on their total test score, so that you have a table of scores for the high-scoring examinees, and a table of scores for the low-scoring examinees. By subtracting the item difficulty for the high-scoring group of examinees from the item difficulty for the low-scoring group of examinees, we have a measure of item

discrimination (use Difficulty2). All of the discrimination values in the example below are positive, indicating items that discriminate positively among high- and low-scoring examinees. Based on this analysis, these items are performing as expected.

Item Discrimination, Twenty-Item Test					
Examinees	Item 1	Item 2	Item 3	Item 4	Total Score*
Low-Scoring Examinees					
3	0	0	0	0	6
2	1	0	1	4	16
1	1	0	3	4	20
10	1	0	1	1	23
High-Scoring Examinees					
5	1	0	3	4	29
7	1	1	3	2	29
8	1	0	4	4	29
4	1	1	4	4	32
9	1	0	4	4	33
11	1	1	3	4	33
12	1	0	4	4	34
6	1	1	4	4	35
Diff2 (High)	1.00	0.50	0.91	0.94	
Diff2 (Low)	0.75	0.00	0.31	0.56	
Difference/ Discrimination	0.25	0.50	0.59	0.38	
Difficulty**	0.92	0.33	2.83	3.25	
Max. Points	1.00	1.00	4.00	4.00	
Difficulty2	0.92	0.33	0.71	0.81	
*Total Score is based on all 20 test items					
**The difficulty values in the white-shaded cells are based on all examinees					

We divide the examinees into the two groups using their Total Score on the assessment (the Total Score is from a test with these four items and another 16 items).

We next compute the Difficulty2 statistic by taking the average item score for each item and dividing by the maximum points for each item.

We finally subtract Difficulty2 for the High group from the Low group to obtain the item discrimination.

The second method involves computing a correlation between the item and the total test score (Excel command = Correl). Using Excel, we computed the discrimination statistics for the first four items on this 20 item assessment. All of these correlations are positive, indicating positive discrimination. All of these items appear to be performing as expected, with respect to discrimination.

Examinees	Item 1	Item 2	Item 3	Item 4	Total Score*
1	1	0	3	4	20
2	1	0	1	4	16
3	0	0	0	0	6
4	1	1	4	4	32
5	1	0	3	4	29
6	1	1	4	4	35
7	1	1	3	2	29
8	1	0	4	4	29
9	1	0	4	4	33
10	1	0	1	1	23
11	1	1	3	4	33
12	1	0	4	4	34
Discrimination	0.74	0.48	0.89	0.64	
*Total Score is based on all 20 test items					

If we found correlations that were close to “0” or were negative, we would want to review the item to be sure that it: a) is clear and not confusing, b) has the correct answer marked in the scoring guide, and c) is representing the same content as the remaining items on the assessment. Test reviewers can then revise or remove items with problematic discrimination statistics.

Identifying Item Bias through Item Discrimination Statistics

Test reviewers can use item discrimination statistics to flag items for having possible bias. In this instance, we define *potential* item bias as an item that has poor discrimination statistics for a sizeable student demographic group (a student demographic group should have at least 20 examinees in it). With this method, we can only identify potential bias. Once potential item bias is identified, the test reviewers can examine the item to see if it has other characteristics associated with item bias.[2]

The process of identifying potential item bias using item discrimination involves computing an item-total test score correlation for each item, as we did using Method #2 in the Item Discrimination section of this document, but this time we compute the correlation for each subgroup and compare the two correlations (subgroup correlation vs. correlation for everyone else). In Excel, this would involve a) sorting the file by the student demographic group (comparing the demographic group to everyone else), then b) computing the item-total test score correlation for each group (the designated demographic group and everyone else). Using this process, reviewers are looking to see if:

- there are any negatively discriminating items for the designated demographic group
- items that have substantially worse discrimination statistics (correlations) for the designated demographic group vs. the total group

Items showing any of the problems described above for the designated demographic group can be reviewed to see if the content of the item contributes to the potential item bias found through this process.

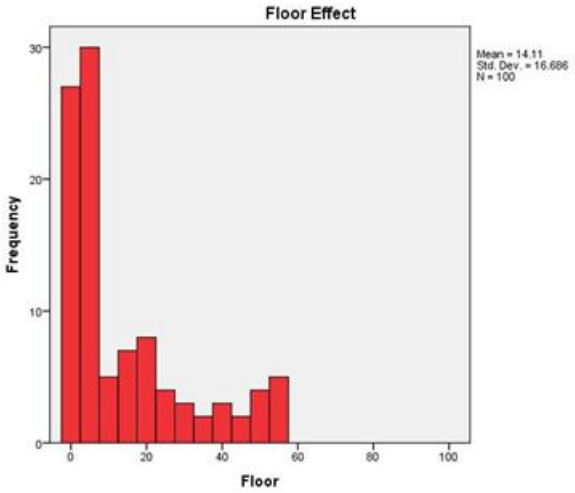
Identifying Floor and Ceiling Effects

When evaluating an assessment for use as a possible DDM, test reviewers are advised to review the distribution of examinee scores to detect any apparent floor and ceiling effects:

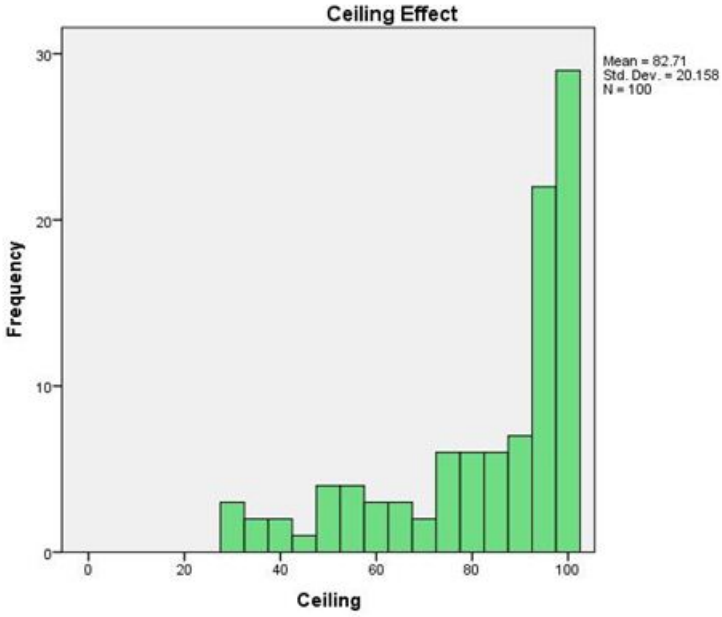
- Floor effects are evident when a substantial number of examinees score at the very low end (or bottom) of the test score scale
- Ceiling effects are evident when a substantial number of examinees score at the highest end (or top) of the test score scale.

Floor and ceiling effects are illustrated in the graphs and frequency distributions below. Floor effects are often found in pre-testing situations, while ceiling effects can be found in post-test situations. Both types of effects should be avoided to ensure that the DDM is measuring examinees as intended. If either floor or ceiling effects are found on a DDM, some of the test items can be revised to either introduce easier items (for floor effects) or harder items (for ceiling effects).

[2] For a list of characteristics associated with item bias, see http://ericae.net/ft/tamu/biaspub2.htm.



Floor Effect	
Score Range	*Frequency*
5	63
10	5
15	6
20	4
25	2
30	4
35	4
40	2
45	6
50	2
55	1
60	0
65	0
70	0
75	0
80	0
85	0
90	0
95	0
100	0



Ceiling Effect	
Score Range	*Frequency*
5	63
10	6
20	6
50	5
15	4
25	4
35	4
40	2
30	2
45	2
55	1
60	0
65	0
70	0
75	0
80	0
85	0
90	0
95	0
100	0