



DYNAMICTM

LEARNING MAPS

Results from External Review during the 2013–2014 Academic Year

Technical Report #14-02

7/25/2014

All rights reserved. Any or all portions of this document may be reproduced and distributed without prior permission provided the source is cited as:

Clark, A., Karvonen, M., & Swinburne Romine, R. (2014). *Results from external review during the 2013–2014 academic year* (Technical Report No. 14-02). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

The present publication was developed under grant 84.373X100001 from the U.S. Department of Education, Office of Special Education Programs. The views expressed herein are solely those of the authors, and no official endorsement by the U.S. Department of Education should be inferred.

Table of Contents

| | |
|--|----|
| Development of DLM’s External Review Process..... | 4 |
| Pilot External Review | 4 |
| On-Site Virtual Review Session..... | 5 |
| External Reviewers..... | 6 |
| External Review Process | 9 |
| Results of Reviews During the 2013–2014 Academic Year..... | 9 |
| Content Team Decisions | 10 |
| English Language Arts Text Review | 13 |
| Conclusion..... | 14 |
| Appendix: Review Criteria for Panelists | 16 |

Results from External Review during the 2013–2014 Academic Year

The purpose of this report is to provide a summary of the outcomes of external review of tasks/testlets and texts reviewed during the 2013–2014 academic year. The report includes a description of the reviewers who were selected to participate in the external review process for the Dynamic Learning Maps™ (DLM) alternate assessment system, a summary of the external review recommendations for individual tasks and testlets reviewed, and a summary of the decisions made by the DLM content teams as a result of the feedback provided by external reviewers.

The external review process comes between two distinct steps in the test development cycle: after an internal review of tasks and testlets, and before the final decision is made regarding which tasks and testlets to field test. Members of three separate review panels conduct external reviews. One panel reviews accessibility, another reviews bias and sensitivity, and a third reviews content. In 2013–14, external review of texts was conducted separately from external review of tasks/testlets. Results for English language arts (ELA) text review are summarized at the end of this report.

A separate review process was used for the initial review of tasks and testlets that were written specifically for blind or visually impaired students. Please see White Paper No. 14-02 for these findings.

Development of DLM’s External Review Process

Pilot External Review

In August of 2013, the first set of external reviews were conducted on-site in Kansas City, Missouri. Reviewers examined content for the fall 2013 pilot assessment and piloted the review criteria used by each panel to evaluate tasks and testlets. A list of the final review criteria is provided in the Appendix. Each state selected two participants for the pilot review session: one for ELA and one for mathematics. A total of 34 participants were included in the pilot review session.

Reviewers were asked to provide feedback on the importance of training topics using a 3-point scale: not at all important, somewhat important, or very important. The majority of reviewers indicated that all topics were very important. The most “very important” responses were recorded for training on the review procedures (97%) and for reviewer responsibilities (92%).

Reviewers were also asked to evaluate the external review experience on a four-point Likert scale, ranging from strongly disagree to strongly agree. All participants either agreed (15%) or strongly agreed (85%) that the process of external review had value as a

professional development experience. All respondents either agreed (27%) or strongly agreed (73%) that there was value in evaluating testlets as a whole in addition to the separate items, and all either agreed (38%) or strongly agreed (62%) that the final decisions made by their table of reviewers reflected his or her own individual opinions about the task/testlet.

Suggestions for improvement were also sought from the reviewers. Suggested areas for improvement included clarifying the accessibility criteria, highlighting key concepts (e.g., linkage levels), and reworking some of the training content both to convey the relative emphasis of topics and to more accurately model the sequential process followed by reviewers.

On-Site Virtual Review Session

Following the pilot of the review process in August of 2013, external review was moved to an online platform. Using refined review processes based on lessons learned from the pilot, both training and review of tasks/testlets were made available online.

To try out the system before it went live consortium-wide, an online tryout was conducted in October of 2014. One district in Missouri was selected for the first round of virtual reviews. Educators within a 50-mile radius of Platte City, MO who volunteered to serve as external reviewers were contacted to participate in an on-site trial run of the virtual review process. DLM staff were present in case of any errors and to collect feedback from the reviewers on areas for improvement; however, reviewers worked independently through training and review of items. The on-site session counted towards each individual’s total external review commitment for the year.

A total of fourteen reviewers participated in the on-site virtual review session. Table 1 describes the number of reviewers by content area.

Table 1

On-Site Reviewers by Content Area

| | English language arts | Mathematics |
|----------------------|--------------------------|-------------|
| Accessibility | 2 | 3 |
| Bias and Sensitivity | 2 | 2 |
| Content | 3 | 2 |
| Total | 7 | 7 |

Participants in the on-site trial also completed a survey on the virtual review process. Of the 14 respondents, 12 (86%) indicated that the overall goals of DLM’s external review process were clear. All 14 respondents reported that they valued the DLM external review process as a professional development experience.

The results of the survey revealed a need to provide a better explanation of what reviewers saw in composite media items, to improve the training videos, and to add clarifying statements in the handbook based on reviewers’ common misconceptions (e.g., a tendency to critique based on features that weren’t relevant to the assigned panel).

External Reviewers

Volunteer reviewers were initially recruited during September of 2013 from DLM partner states to review testlets (sets of 3–5 items) for the DLM alternate assessment system, to be administered during the 2013–2014 academic year. A recruitment letter was prepared by DLM staff and disseminated to state partners for distribution. The letter included a brief overview of the external review process and asked interested individuals to complete an online Qualtrics survey to provide their qualifications and contact information.

DLM received information from a total of 932 volunteer reviewers during the 2013–2014 academic year. Reviewers included classroom teachers and other local educators, representatives from state departments of education, and university faculty members. The average number of years of experience in preschool–12th-grade education was 16 years. Approximately 65% of volunteers had taught students with significant cognitive disabilities within the last ten years. A total of 34% indicated that they had previous review experience for another large-scale assessment. Table 2 describes the number of volunteer reviewers by state.

Table 2

Number of Reviewers by State

| State | Reviewers | |
|----------------|------------------|---------------------|
| Alaska | 51 | 5.5 ⁰ % |
| Colorado | 1 | 0.1 ⁰ % |
| Illinois | 46 | 4.9 ⁰ % |
| Iowa | 7 | 0.8 ⁰ % |
| Kansas | 153 | 16.4 ⁰ % |
| Michigan | 36 | 3.9 ⁰ % |
| Mississippi | 79 | 8.5 ⁰ % |
| Missouri | 177 | 19.0 ⁰ % |
| New Jersey | 110 | 11.8 ⁰ % |
| North Carolina | 10 | 1.1 ⁰ % |
| Oklahoma | 60 | 6.4 ⁰ % |
| Pennsylvania | 1 | 0.1 ⁰ % |
| Utah | 29 | 3.1 ⁰ % |
| Vermont | 69 | 7.4 ⁰ % |
| Virginia | 2 | 0.2 ⁰ % |
| Washington | 2 | 0.2 ⁰ % |
| West Virginia | 39 | 4.2 ⁰ % |
| Wisconsin | 48 | 5.2 ⁰ % |
| Other* | 12 | 1.2 ⁰ % |
| Total | 932 | |

**From non-DLM states. Since external review recruitment was done through DLM state partners, it was assumed that these reviewers were intentionally recruited.*

Of the original 932 volunteer reviewers, a total of 414 did not continue with the review process. Reasons for discontinuation included not completing the required training, not providing qualifying information, or volunteering too late to be placed on a complete panel. This reflects an attrition rate of 44%.

The remaining 518 reviewers were assigned to one of three types of review panels: accessibility, bias and sensitivity, or content review. In order for reviewers to qualify for the accessibility review panel, they either had to have at least one year of experience (preschool through grade 12) working with students with significant cognitive disabilities or have at least one year of experience working with alternate assessments based on alternate achievement standards. Priority for the bias and sensitivity review panel was given to reviewers who selected a race other than white, but other reviewers were included in the panel as needed. To qualify for the content review panel, reviewers had to have at least one year of educational experience (preschool through grade 12) in ELA or

mathematics. Regardless of panel type, all reviewers had to have access to a secure computer and agree to complete at least two rounds of review.

A total of 155 reviewers were assigned to an accessibility review panel, 107 were assigned to a bias and sensitivity review panel, and 251 were assigned to a content review panel. There were fewer bias and sensitivity reviewers because they were able to complete more reviews in less time compared with reviewers in other areas. Likewise, the content panel had the most review criteria to evaluate, so their review time per testlet was longer than that of the other two review panel types.

Of the 518 reviewers assigned to a panel, only 357 completed two or more review assignments, reflecting an attrition rate of 62% of the original 932 volunteer reviewers and a 31% attrition rate at this stage. Only 215 reviewers completed all assigned reviews. Because of the high attrition rate and the need to complete reviews in time for field testing, a small group of reviewers, called “Power Reviewers,” were selected to review for all three panel types. A total of 24 Power Reviewers were selected from the pool of 518 volunteer reviewers based on the quality of their previous work. They then completed reviews on a more intensive schedule. They also received an honorarium based on the volume of testlets they reviewed. In addition, a total of five part-time paid reviewers were selected to increase the pool of reviewers who would complete assignments on time and with a high standard for quality reviews. These paid reviewers met the qualifications for all three panel types and completed the training for all three panels.

The number of reviewers by content area are presented in Table 3 below.

Table 3

Number of Reviewers by Content Area

| Review Panel | English | |
|----------------------|---------------|-------------|
| | Language Arts | Mathematics |
| Accessibility | 57 | 91 |
| Bias and Sensitivity | 46 | 55 |
| Content | 104 | 136 |
| Hourly Reviewer | 2 | 3 |
| Power Reviewer | 13 | 11 |
| Total | 222 | 296 |

External Review Process

Prior to receiving their first external review assignment, volunteer reviewers had to complete a set of training activities in order to become familiar with the DLM external review process. The first training module included a general overview of the process of external review. Following completion of the first module, reviewers submitted a test security agreement. Next, reviewers completed a module that was specific to the type of reviews they had been assigned (i.e., accessibility, bias and sensitivity, or content). A final module consisted of the procedures for completing reviews. After completion of all three modules, reviewers completed a quiz and a practice activity to familiarize themselves with the review process. Volunteers who completed the security agreement, quiz, and practice activity were then assigned testlets to review. Those who did not complete all three requirements, including a minimum score on the quiz, were not given content to review. The quiz allowed multiple opportunities to achieve a passing score.

External reviewers virtually reviewed tasks/testlets by previewing them on their own computer and completing an online survey about each set of items. During the process of reviewing each task/testlet, reviewers asked themselves the following question: “Does this task/testlet meet minimal standards for acceptability based on my panel’s criteria?” *Acceptability* was defined as meeting minimum standards for field testing readiness. Based on their response to this question, reviewers made one of three recommendations:

1. Accept: the task/testlet is within acceptable limits for field testing
2. Revise: the task/testlet violates one or more of the review criteria; however, the task/testlet has potential merits and can be acceptable for field testing after revisions to address the criteria
3. Reject: the content of the task/testlet is fundamentally flawed; revision would not bring the task/testlet to acceptable limits

If an “accept” recommendation was made for the task/testlet, no further information was needed from the reviewer. If a recommendation of revision was made, the reviewer’s comment identified the problem as well as a proposed solution. For a recommendation of rejection, the reviewer’s comment identified the problem only.

Results of Reviews during the 2013–2014 Academic Year

The majority of the content reviewed during the 2013–2014 academic year was included in the fall pilot and spring field testing events. On a limited basis, content for the upcoming 2014–2015 school year was also reviewed. The reviewers’ ratings for tasks and testlets are presented in Tables 4 and 5 below. Because multiple reviewers reviewed

each task and testlet, these values represent the count of recommendations rather than the number of unique tasks and testlets reviewed.

For ELA, the percent of tasks or testlets rated as “accept” ranged from 72% to 91%. The rate at which content was recommended for rejection ranged from 1% to 5%.

Table 4

ELA External Review Recommendations

| | Accessibility | | Task Bias | | Content | | Access | | Testlet Bias | | Content | |
|--------------|---------------|------|-------------|------|-------------|------|------------|-----|--------------|-----|------------|-----|
| | Accept | 2010 | 79% | 2351 | 91% | 2141 | 72% | 625 | 80% | 738 | 91% | 691 |
| Revise | 455 | 18% | 204 | 8% | 678 | 23% | 142 | 18% | 67 | 8% | 184 | 20% |
| Reject | 71 | 3% | 30 | 1% | 138 | 5% | 14 | 2% | 5 | 1% | 33 | 4% |
| Total | 2536 | | 2585 | | 2957 | | 781 | | 810 | | 908 | |

For mathematics, the percent of tasks or testlets rated as “accept” ranged from 76% to 88%. The rate at which content was recommended for rejection ranged from 2% to 3%.

Table 5

Mathematics External Review Recommendations

| | Accessibility | | Task Bias | | Content | | Access | | Testlet Bias | | Content | |
|--------------|---------------|------|-------------|------|-------------|------|-------------|------|--------------|------|-------------|------|
| | Accept | 3957 | 76% | 3879 | 88% | 4536 | 78% | 1090 | 76% | 1087 | 88% | 1257 |
| Revise | 1103 | 21% | 410 | 9% | 1128 | 19% | 312 | 22% | 122 | 10% | 317 | 20% |
| Reject | 139 | 3% | 125 | 3% | 165 | 3% | 36 | 3% | 27 | 2% | 31 | 2% |
| Total | 5199 | | 4414 | | 5829 | | 1438 | | 1236 | | 1605 | |

Content Team Decisions

Because multiple reviewers examined each task and testlet, external review ratings had to be compiled across panel types. Content teams reviewed and summarized the recommendations provided by the external reviewers for each task and testlet. Decision options were broken into five categories:

1. No pattern of similar concerns; accept as-is
2. Pattern of minor concerns, will be addressed
3. Major revision needed

4. Reject
5. More information needed

Content teams documented the decision category for each task and testlet, as well as the reason for a decision of 2 through 5, using the codes provided in the Appendix for each of the three panel types.

Following this process, content teams made a final decision to accept, revise, or reject each of the tasks and testlets. The number of accept, revise, and reject decisions made by each content team are included in Tables 6 and 7.

Table 6

ELA Decisions Based on External Review

| | Task | | Testlet | |
|--------------|-------------|-----|------------|-----|
| Accept | 1881 | 90% | 485 | 78% |
| Revise | 165 | 8% | 128 | 20% |
| Reject | 49 | 2% | 13 | 2% |
| Total | 2095 | | 626 | |

The ELA content team retained 98% of tasks and testlets sent out for external review. Of the tasks and testlets that were revised, most required only minor changes (e.g., minor rewording but concept remains unchanged), as opposed to major changes (e.g., stem or option replaced). The ELA team made a total of 124 minor revisions to tasks and 84 minor revisions to testlets.

Of the 165 ELA *tasks* that were revised, 134 (81%) were flagged for having a content issue. A total of 23 (14%) were flagged for an accessibility issue. Only 1 task (1%) was flagged for a bias and sensitivity issue. A total of 7 tasks (4%) were flagged for having an issue in more than one area (i.e., accessibility, bias and sensitivity, and/or content).

Of the 49 ELA *tasks* that were rejected, 21 (43%) were flagged for having a content issue, 3 (6%) were flagged for an accessibility issue, and 21 (43%) were flagged for having a bias and sensitivity issue. A total of 4 (8%) were flagged for having an issue in more than one area.

Of the 128 ELA *testlets* that were revised, 105 (82%) were flagged for having a content issue, 12 (9%) were flagged for an accessibility issue, and 2 (2%) were flagged for having a bias and sensitivity issue. A total of 9 (7%) were flagged for having an issue in more than one area.

Of the 13 ELA *testlets* that were rejected, 4 (31%) were rejected for having a content issue, 1 (8%) was rejected for an accessibility issue, and 6 (46%) were rejected for having a bias and sensitivity issue. A total of 2 (15%) were flagged for having an issue in more than one area.

Table 7

Mathematics Decisions Based on External Review

| | Task | | Testlet | |
|--------------|-------------|-----|-------------|-----|
| Accept | 3391 | 88% | 878 | 79% |
| Revise | 453 | 11% | 226 | 20% |
| Reject | 17 | 1% | 8 | 1% |
| Total | 3861 | | 1112 | |

The mathematics content team retained 99% of tasks and testlets sent out for external review. As with ELA, most revisions made to tasks and testlets were minor. The mathematics team made a total of 387 minor revisions to tasks and 186 minor revisions to testlets.

Of the 453 mathematics *tasks* that were revised, 205 (45%) were flagged for a content issue, 131 (29%) were flagged for an accessibility issue, and 11 (2%) were flagged for having a bias and sensitivity issue. A total of 106 tasks (23%) were flagged for having an issue in more than one area.

Of the 17 mathematics *tasks* that were rejected, 3 (18%) were rejected for a content issue, 3 (18%) were rejected for an accessibility issue, and 5 (29%) were rejected for having a bias and sensitivity issue. A total of 6 (35%) were flagged for having an issue in more than one area.

Of the 226 mathematics *testlets* that were revised, 105 (46%) were flagged for having a content issue, 52 (23%) were flagged for an accessibility issue, and 2 (1%) were flagged for having a bias and sensitivity issue. A total of 67 tasks (30%) were flagged for having an issue in more than one area.

Of the 8 mathematics *testlets* that were rejected, 1 (12.5%) was rejected for a content issue, 1 (12.5%) was rejected for an accessibility issue, and 2 (25%) were rejected for having a bias and sensitivity issue. A total of 4 (50%) were flagged for having an issue in more than one area.

English Language Arts Text Review

As a part of the external review process, reviewers also evaluated the DLM ELA texts. Reviews were conducted between November of 2013 and April of 2014.

External review of texts was conducted separately from external review of testlets. Each state was asked to recruit one reviewer to divide the review responsibilities so that there would be one evaluation per text, per state. DLM solicited reviewers who had experience reviewing ELA content and/or experience with students with significant cognitive disabilities.

The first group of reviewed texts ($N = 114$) was scheduled for potential use in Field Tests #1 and #2. State members (or their designees) were invited to participate in a review of all DLM texts that were under consideration for the first field test. Reviewers were provided with a short summary of how DLM ELA texts are created and used, as well as a description of the internal guidelines used for text development. Copies of all texts were provided to state members for review. Reviewers submitted feedback using an online survey. Reviewers were asked to identify issues regarding bias and sensitivity in ELA stories and informational texts.

The pool of texts that were available for potential use in Field Test #3 or subsequent field tests was larger ($N = 181$). In December of 2013, this pool was divided into three batches with consecutive review windows. Reviewers completed asynchronous, online training activities that included reading a description of the ELA text construction process, a short video introduction to ELA texts, and a review of the criteria used in text development. All texts were provided via a secure file-sharing platform, and the descriptions of how texts were assigned to nodes were included in a spreadsheet for each review batch. Reviewers evaluated texts using criteria related to content, accessibility, and bias and sensitivity. Reviewers were asked to identify texts that did not meet the text development guidelines. When a text was judged by a reviewer to not meet the criteria, an online survey tool was used to collect feedback about the text. Reviewers selected criteria related to their feedback, entered free-response comments about the problem, and suggested possible revisions. During the first two windows, the criteria and comment sections were identical to the first group of texts. For the third window, evaluators also had an opportunity to recommend outright rejection of the text.

After each review window closed, DLM staff evaluated the collective set of recommendations for each text. In cases where reviewers identified problems, texts were revised to enhance language clarity, to improve the logical presentation of content, and to remove bias or sensitivity issues with photos. When multiple reviewers raised the same issue regarding sensitive content in a text, the text was revised, where possible, to

eliminate the problem. In cases where revision would compromise the utility of the text to measure its assigned node(s), the text was rejected.

A summary of reviews of ELA texts is presented in Table 8.

Table 8

Summary of Reviews and Results for Each Group of Texts

| | FT 1-2 | Batch A | Batch B | Batch C |
|---|----------|----------|----------|----------|
| Number of states submitting reviews | 7 | 10 | 10 | 10 |
| Total texts available for review | 114 | 46 | 74 | 61 |
| Number of texts receiving a comment | 94 | 42 | 42 | 56 |
| Total comments received | 183 | 158 | 240 | 182 |
| Number (percent) of texts recommended for rejection | | | | 17 (28%) |
| Number (percent) of texts revised based on reviews | 16 (14%) | 33 (72%) | 25 (34%) | 21 (34%) |
| Number (percent) of texts rejected based on reviews | 1 (1%) | 2 (4%) | 0 (0%) | 10 (16%) |

Note: FT = field test; FT 1-2, Batch A, and Batch B did not give reviewers the option to recommend rejection of texts.

Conclusion

The vast majority of content externally reviewed during the 2013–2014 academic year was either accepted outright or with revisions. Only 1% of mathematics tasks and testlets and only 2% of ELA tasks and testlets were rejected. This finding provides support for DLM’s approach to evidence-centered design, whereby tasks and testlets are specifically created using Essential Element Concept Maps with accessibility, bias and sensitivity, and content in mind.

Recruitment for external review for the 2014–2015 academic year is underway. Several changes have been made for the next round of external review based on outcomes from the initial year. First, recruitment timelines were made more specific. Rather than having a single ongoing volunteer window, three phases of volunteering are being implemented. Each phase has a deadline by which to submit the volunteer survey, complete the required training, and receive the first assignment. In addition, the opening of the three phases is being staggered across the months of August and September to account for varying start times across the states in the consortium.

Another area that is changing significantly as a result of outcomes from the 2013–2014 session is the criteria for evaluation of bias and sensitivity in testlets and ELA passages. A series of discussions with state partners led to the development of revised

criteria for bias, sensitivity, and the use of images in testlets. These changes have been incorporated into training materials and will be implemented in external reviews during the 2014–2015 academic year.

Because of the high attrition rate observed during the 2013–2014 academic year, external reviewers will be compensated for completing sets of testlet reviews during the 2014–2015 academic year. As an additional incentive, compensation amounts will increase based on the number of collections completed by the reviewer. In an effort to track content more easily, reviewers will only be assigned a single collection at a time. In addition, collections will be of slightly larger size to allow for more content to be reviewed at a single time.

Review of testlets for students who are blind or visually impaired will also be conducted during the 2014–2015 academic year. Priority review of testlets for students who are blind or visually impaired will be given to volunteers who either have experience working with students with significant cognitive disabilities or experience working with students who are blind or have low vision.

A final change being made to the external review process results from the two assessment models being implemented in the 2014–2015 academic year: the integrated model and the year-end blueprint model. Although panel types and qualifications for placement on each panel will remain the same, the reviewer’s state will now serve as an additional piece of information informing panel assignment, with reviewers from year-end model states reviewing multiple Essential Element testlets and reviewers from the integrated model states reviewing single Essential Element testlets (which include instructionally embedded assessments).

Appendix: Review Criteria for Panelists

CRITERIA FOR ACCESSIBILITY PANELS

Tasks

1. The text within the item provides an appropriate level of challenge and maintains a link to grade-level content without introducing unnecessary, confusing, or distracting verbiage. The text uses clear language and minimizes the need for inferences and prior knowledge to comprehend the content.
2. Graphics are clear and do not introduce confusion. Graphics can be presented in tactile form.

Testlets

3. The testlet is instructionally relevant to students for whom it was written and is grade-level appropriate.
4. The testlet does not introduce barriers for students with (a) limited working memory, (b) communication disorders dependent on spoken English grammatical structures, or (c) limited implicit understandings of others' emotions and intentions.

CRITERIA FOR BIAS & SENSITIVITY PANELS

Tasks (all bias criteria)

1. Item does not require background knowledge outside the bounds of the targeted content.
2. There is a fair representation of diversity in ethnicity, gender, disability, and family composition.
3. Stereotypes are avoided. Appropriate labels are used for groups of people. People-first language is used for individuals with disabilities.
4. Language used does not prevent nor advantage any group from demonstrating what they know about the measurement target.

Testlets (sensitivity criterion)

5. Testlet is free of content that is controversial, disturbing, or emotionally charged due to issues of culture, region, gender, religion, ethnicity, socioeconomic status, occupation, or current events.

CRITERIA FOR CONTENT PANELS

Tasks

1. The item assesses the content of the targeted node.
2. The level of depth of knowledge required in the node matches the depth of knowledge identified for the item.
3. The content of the item is technically correct (wording and graphics).
4. Item answer options contain only one correct answer (the key), distractors are incorrect and not misleading, and nothing in the item cues the correct response.

5. The item type is logical and appropriate for the content being assessed, and the graphics contribute to the quality of the item.

Testlets

6. The testlet is instructionally relevant to the students for whom it was written and is grade-level appropriate.
7. Embedded items are placed within the story text at logical places, and conclusion items are placed at the end (ELA only).