eOLA
OUT-OF-SCHOOL YOUTH LITERACY ASSESSMENT

Expert Review of EDC's Out-of-school youth Literacy Assessment (OLA)

August 2013

EDC Learning transforms lives.

Expert Review of EDC's Out-of-school Literacy Assessment (OLA)

August 2013

FOR MORE INFORMATION, PLEASE CONTACT:

eOLA@edc.org | www.eOLA.edc.org

Produced by Education Development Center, Inc.
43 Foundry Avenue Waltham, MA 02453-8313, USA
Boston - Chicago - New York - Washington, DC

Summary

In August 2013, Education Development Center, Inc. (EDC) gathered experts in youth literacy, linguistics, and international assessment and evaluation to review EDC's Out-of-school youth Literacy Assessment (OLA) instrument and accompanying data. The purpose of the review was to:

1) Evaluate the reliability and strength of the existing reading assessment and protocols.
2) Provide recommendations for improving the assessment.

Overall, the panel found that OLA is a strong measure of early reading skills for older populations, and it offers valuable feedback on young adult literacy program effectiveness. Two of the primary strengths identified by the panel are the functional literacy sub-test, which measures real-life reading capacities, and the extensive background questions that cover educational, social, linguistic, and economic information about the respondent. The panel's recommendations focused on expansion of existing subtests, including adding more items to the functional literacy section, and developing an additional sub-test that measures oral language comprehension.

OLA Overview

International efforts to promote literacy development for young adult readers are increasing, yet tools for assessing basic literacy lag behind. In order to respond to this need, EDC developed OLA to measure literacy skills of older youth and young adults, particularly those who are living in extreme poverty or post-conflict environments with minimal literacy acquisition. OLA builds on reading research and best practices in evidence-based adult literacy instruction and assessment and incorporates real-life reading items that measure the foundational literacy skills that youth and adults may already have, such as locating words on food labels or reading signs. In addition to real-life reading, OLA orally assesses letter naming, letter sounds/syllables, word list reading, short passages reading and comprehension. The instrument also includes a background section that captures relevant demographics. An assessment of basic writing skills is being developed to accompany the reading section of OLA.

OLA is orally administered one-one-one with each reading test taking about 15-20 minutes per respondent. Each respondent is given a stimulus with text and pictures that correspond to each of the subtests. The administrator records the responses either on paper or electronically. The electronic version of OLA (eOLA), available for use on laptops or tablets, leads to more efficient data collection and analysis because it allows for real-time data

management and monitoring of data quality and eliminates the costs and errors associated with data entry.

OLA was primarily developed to provide summative information for youth projects; that is, to provide critical data on basic reading skills to evaluate the project's impact. OLA results also inform implementation, such as appropriateness of instructional approaches, curriculum, and teaching and learning materials. OLA's demographic section provides valuable information on a population for which only minimal literacy data often exist in national censuses and surveys. The demographics also allow sub-analyses by gender, age, urban/rural residence, and other characteristics. In addition to being used by projects, OLA can also be administered to a general youth population to provide information on literacy that will be useful to government institutions, development organizations and other stakeholders.

Data have been collected in Liberia (in English), Rwanda (in Kinyarwanda) and in Mali (in Bambara and Songhai). In Liberia, the data have helped technical teams improve the alternative basic education intervention, while the OLA training of administrators has raised awareness of Ministry of Education officials about differences that exist between assessing reading skills of young children and those of older youth. In Mali, data are being utilized by curriculum supervisors to ensure that reading skill categories are being addressed in the two target languages. In Rwanda, data are being used to inform project management and technical programming for youth reading at a grade equivalency of 4-6.

Expert Panel and Review Process

EDC convened a panel of external experts to review the quality and content of OLA and to make recommendations for improvements and future use. The panel, which met in Washington, DC for two days in late August 2013, was composed of:

- Carolyn Adger, Center for Applied Linguistics
- Manuel Cardoso, UNESCO Literacy Assessment and Monitoring Program (LAMP)
- Mary Beth Curtis, Lesley University
- Jeff Davis, consultant and psychometrician
- Irwin Kirsch, Center for Global Assessment, Education Testing Service
- Cristine Smith Crispin, Center for International Education, University of Massachusetts, Amherst

EDC reading specialist Nancy Clark-Chiarelli joined the panel as an ex-officio member. The meeting was also attended by John Comings, USAID Education Policy Advisor, and EDC staff members Brenda Bell, Alejandra Bonifaz, Craig Hoyle, Emily Morris, and John Strucker.

In advance of the meeting, panel members were given copies of the OLA instrument and of the administration guides for three languages (English, Kinyarwanda, and Bambara); Comparative Analysis of the OLA and Early Grade Reading Assessment (EGRA) for the USAID Advancing Youth Project, Liberia (2012); Baseline Data Report, USAID Advancing Youth Project, Liberia (2012); and a list of key review questions.

The panel's deliberations were structured to give panelists sufficient information to make informed judgments, with the flexibility to set their own agenda. On the first day, panelists were given a thorough orientation to the OLA, using the English version as administered in Liberia as the case study. Topics included the rationale and development process for OLA; detailed descriptions of the context in which OLA is administered, including information on learners, teachers and classroom environments; procedures for training test administrators and field administration; detailed descriptions of each of the OLA's sub-tests; an overview of the reliability and quality control measures for the instrument; and the procedures for adapting OLA to different countries and languages. This overview was followed by a detailed psychometrics presentation on data from the Liberia's pre and post-tests in English, including equating of the pre and post-tests. Data from the OLA and EGRA comparative analysis were also reviewed.

During these initial sessions, panelists exchanged ideas and offered preliminary suggestions. Equipped with background information and the initial discussions around key issues, the panel spent the remaining 1 ½ days addressing EDC's specific questions about OLA and, where appropriate, making recommendations for modifications and improvements. For most of the sessions the panel functioned as a committee of the whole, but at times members met briefly in smaller break-out groups (e.g., psychometricians, literacy survey experts, and reading and language experts).

Background/Demographic Section

The following summary of the panel's findings is drawn from the report of the panel's deliberations prepared by panelist Jeff Davis and the written responses submitted by each panelist at the end of the meeting. As noted earlier, the panel focused on the English version of the OLA and the data sets from administration in Liberia with youth enrolled in the USAID Advancing Youth Program. This summary is organized by the key findings related to the sub-tests and the overall assessment.

Below are findings and recommendations specific to selected OLA sections and sub-tests:

1. **Background/Demographic Section**
   FEEDBACK The extensive background questionnaire provides critical information for project teams and ministry officials about program participants, including their economic,

social, educational, and linguistic backgrounds. It also gives analysts the ability to look at skill development by different sub-groups of the tested population.

RECOMMENDATIONS

- Add more demographic questions on language background, including language used for work/business and language in which the students first learned to read or write (formally or informally).
- Add questions to capture respondent's access to print and text in their current home and/or work environments.

2. **Letter Sound and Syllable Sub-test**

FEEDBACK Upon review of the Liberia data, the panel verified that this section was difficult for the respondents given that letter sounds have not been consistently taught in previous literacy classes there. They noted that this might also be the case for a number of other countries. They agreed with the adaptation of this section that was made for Rwanda: given the transparent orthography of Kinyarwanda and how literacy is taught in that language, it made sense to replace letter sounds with common syllables.

RECOMMENDATIONS

- Noting the difficulty of the letter sound subtests for many respondents, the panel recommended moving that test to the end so as not to unduly discourage respondents at the beginning of the test.
- They also recommended that as OLA is adapted for different languages and orthographies in the future, new phonemic awareness sub-tests appropriate for those languages and orthographies should be developed.

3. **Real-Life Reading Sub-test**

FEEDBACK There was consensus among panelists that the real-world items were essential for understanding the reading development of youth and adults. The panel stressed that the real-world items bridge the gap between a learner's developing abilities in the components of reading (e.g., word reading and fluency) and her/his ability to use literacy in the contexts of work, family health, and civic participation. In addition, the panel felt that OLA's real-world items can help ministries of education and labor assess adult learners' readiness for participation in various sectors of the economy.

RECOMMENDATION

- The panel recommended that the number of real-world items (currently five) be increased to include additional beginning-level items.

4. **Word Reading Sub-test**

FEEDBACK There were two main findings: 1) the panel approved of OLA's method for constructing word reading tests based on sampling words from curriculum materials.

However, panel members argued persuasively that OLA's 100-word test, which covers a whole page in one large block of words, might be intimidating for many beginning readers. 2) Although USAID and other donors focus on oral reading fluency with words-correct-per-minute (wcpm) as the main metric, the panel stressed that oral reading accuracy should be given more weight than rate, especially when interpreting results for beginning readers. However, test administrators must be trained to accept dialect differences in word pronunciation as correct.

RECOMMENDATIONS

- The word reading test should be made less intimidating by shortening the test from 100 words to 45 words and arranging the words into three separate 15-word lists corresponding roughly to easy, medium, and more difficult words.
- When interpreting and reporting word reading results, accuracy (the number of words read correctly) should be emphasized over rate (number of words-correct-per-minute) for beginning level learners.

5. **Oral Reading Passages and Comprehension Sub-test**

FEEDBACK As with the word reading tests, the panel agreed that the OLA oral reading passages can be used to measure rate (wcpm). However, as with their comments in regard to Word Reading, the panel felt that accuracy should be emphasized over rate for beginning level readers. The panel also had some feedback regarding the comprehension questions asked orally after each passage is read. Similar to EGRA, OLA asks comprehension questions after each passage is read aloud. Some panel members felt strongly that comprehension questions should not be asked following oral reading passages, especially of adults who may be so focused on accurate decoding that they do not attend closely to meaning. Panelists confirmed the importance of the OLA procedures that allow learners to look back at the text when answering comprehension questions, as this corresponds to how adults often interact with print.

RECOMMENDATIONS

- The panel suggested that in interpreting and reporting oral reading passage results, accuracy (the number of words read correctly) should be emphasized over rate (number of words-correct-per-minute) for beginning level learners.
- To address their concerns about asking comprehension questions following oral reading, the panel suggested that questions might be asked on one of the two oral reading passages.

6. **Silent Reading Passage and Comprehension Sub-test**

FEEDBACK Based on the data presented, the panel noted that test takers appeared to perform more consistently on informational passages than they did on fictional passages. They further noted that informational reading, like real-life reading, is central to how adults use literacy in their work, family life, and civic participation. Secondly, the panel members suggested possible revisions for some comprehension questions which, based on statistical analyses, appear to perform poorly.

RECOMMENDATIONS

- Fictional passages should be eliminated in favor of informational passages.
- Continue to use IRT and other statistical techniques to revise poorly performing questions and answer choices.

The following findings pertain to the full instrument:

7. **Inferences about early reading skills of the target population**

FEEDBACK The review panel determined that OLA provides sufficient information to make basic inferences about early reading skills, when it is administered to respondents who are proficient in the language of the text. More reliable inferences could be made by adding an oral proficiency measure. A review of the data indicated that some of the sub-tests should be reordered, based on familiarity and ease of answering, to make test-takers more comfortable.

RECOMMENDATIONS

- Include an assessment of oral language comprehension.
- Add additional subtests of oral pre-reading skills, such as initial sound recognition and blending.
- To address the very low skill levels and lack of test-taking experience of the target population, move the letter sound sub-tests to last and the real-life reading sub-test first.
- Break the 100 word reading sub-test into three shorter and progressively more difficult word lists.

8. **Ability of OLA to inform policy decisions and program design and improvement**

FEEDBACK Panelists agreed that OLA data would be useful to policy makers and donors for program evaluation purposes, including instructional effectiveness and amount and rates of growth in reading skills. They observed that the ultimate validity of OLA to inform policy will depend on its direct connection to literacy curriculum and learner performance standards, variables that are dependent on the country-by-country development of specific curriculum, standards, and benchmarks.

RECOMMENDATIONS

- The panel supported the idea of developing cut-points for the following sub-tests: letter naming, letter sounds, word reading, oral reading passages, and silent reading comprehension. The panel added that the starting place for establishing cut-points would be natural breaks in a country's data set informed by the demographic data, such as the three groups observed in the Liberian learner population: non-readers, emergent readers, and beginning readers.

- In order to establish cut-points for the real life literacy subtest, additional items would be needed to provide an adequate distribution of learners from least skilled to highest skilled.

9. **Process for adapting OLA to different contexts and languages**

FEEDBACK Overall, the panel determined that for alphabetic languages the OLA subtests for word reading, letter naming, and letter sounds can be expected to function well. However, in adaptations for syllabic languages, letter naming and letter sound tasks would not be relevant. The panel also noted that real-world items might tend to vary quite a bit from country to country, depending on what kind of environmental print is present.

RECOMMENDATIONS

- When adapting OLA for languages with transparent orthographies like Spanish or Hindi, it may make sense to skip letter sounds and go directly to assessing common syllables or onset-and-rime letter combinations.

10. **Psychometric Procedures for Equating the Pre and Post Tests (Forms A and B)**

FEEDBACK The panel felt that EDC's memo on procedures for equating pre and post tests was clear. They did suggest issues to consider when equating, including the continued need to collect pilot data for each adaptation of OLA to ensure comparability of forms in new administrations; an exploration of equating methodologies such as means equating, linear equating, or methods based upon item-response theory; and conducting additional item-level analyses such as differential item functioning (DIF) on both pilot and primary samples to ensure items are functioning as intended.

The panel acknowledged the challenges with standardizing OLA pilot procedures across administrations and contexts, and stressed the need to conduct pilot testing and to allow time for revision based upon results. Even in situations where a large-scale pilot might not be feasible, the panel felt that a pilot where subjects received comparable sections from forms A and B would provide empirical evidence against which to balance, and potentially refine, the judgment of subject matter experts about the equivalence of pre and post forms.

To examine the equating methods, the panel was presented with item-level, section-level, and total test scores from forms A and B for a group of Level 1 students. The panel found IRT to be useful for looking at large data sets and for providing information when looking at OLA as a whole. However, given the large sample size requirements and strict unidimensionality assumptions of IRT, and the likely-smaller pilot samples in most OLA use contexts, the panel favored the use of classical item statistics and examining equating on a section by section basis.

RECOMMENDATIONS

- The panel recommended using classical statistics (such as percent of students answering correctly) and reliability (where applicable) to equate related sections. The

panel felt that this was important, as individual section scores would likely provide more information from a policy standpoint than reporting a total OLA score across sections.

11. Psychometric Analysis of Data using Item Response Theory (IRT) and Classical Analysis

FEEDBACK Overall the panel felt that EDC's overall data analysis and approach was sound, with a place for additional analyses. The panel was presented with results based on IRT and classical item analysis for Level 1 and Level 2 students in Liberia. Results for the analysis of the test as a whole using both IRT and classical item analysis showed that the test was extremely difficult for Level 1 subjects, but better suited for Level 2. Panel members felt that the data would be of interest to policy makers particularly because so little information for this population of students currently exists.

While achievement was low for the tested population on basic tasks such as letter naming and word reading, the panel noted that the reliability for these tasks was high (>.90). For the real-life reading and silent comprehension sections, however, reliabilities were much lower. The panel felt that this was both due to the difficulty of these tasks and to the low number of items in each of these sections. While no data were available for Level 3 subjects, the panel believed that the silent reading section would be most applicable to that population.

RECOMMENDATIONS

- The panel recommended the inclusion of additional items in the real-life section and reading comprehension section to improve reliability as well as the sensitivity of the instruments for detecting growth.
- In addition to the analyses conducted by the OLA team and planned differential item functioning (DIF) analyses during piloting, the panel suggested that Latent Class Analysis (LCA) or Latent Transition Analysis (LTA) may be useful for identifying instructionally relevant sub-groups of learners based on their profiles of reading sub-skills and demographic characteristics.

12. Training and Monitoring of Data Collection

FEEDBACK The panel reviewed EDC's assessment training plan template and materials and felt they were thorough, with the key training areas incorporated. They reiterated the importance of measuring administrator inter-rater reliability (IRR) during the training, and were happy to see that there was an IRR test as well as a number of IRR activities built into the training design.

RECOMMENDATIONS

- Continue to follow the training plan and ensure that administrators implement best practices in working with low-literate youth to make them comfortable during the testing.

Summary of Key Recommendations

The result of the panel's deliberations was an overall positive evaluation of OLA accompanied by specific feedback recommendations for strengthening the tool. These recommendations will inform the on-going administrations of the OLA, and in the long term will ensure the continued reliability and accuracy of the tool. Below are the recommendations organized into short-term (prior to the next OLA administration) and long-term (as the resources are available).

SHORT-TERM	LONG-TERM
• **Sub-tests order.** Ease the testing burden on very beginning adult learners by changing the order of several sub-tests.	• **Oral language sub-test.** Develop an oral language sub-test
• **Demographic section.** Include questions about language and technology use	• **Phonemic awareness.** Create a bank of early reading sub-tests to measure phonemic awareness
• **Real-life reading.** Create additional items at the beginning level.	• **Silent-reading sub-test.** Develop passages that are linked to country-specific curriculum
• **Word reading.** Create three leveled short subtests out of the current 100-word list	• **Cut-points.** Develop a procedure for establishing OLA cut-points
• **Comprehension questions (oral reading and silent reading).** Address the questions and answer choices that did not discriminate well among respondents	• **Analyses.** Use IRT and LCA/LTA to produce more finely-tuned and broadly useful analyses of OLA data

Appendix 1: Panel's Guiding Questions

1. **Inferences about early reading skills**
   Given the limitations in administration time and the skills of the administrators, does OLA provide enough information to make inferences about the population's early reading skills? (i.e., via letter names, letter sounds, word reading, and oral reading.) For example, it might be nice to have blending and pseudo-words, but these skills may be too difficult to test accurately under field conditions.

   Are there any other relatively easy-to-administer sub tests that we could add?

2. **Other inferences**
   To what extent does the instrument, in its current form, support inferences about:
   - Individual growth
   - Regional growth
   - Policy level decisions
   - Targeting interventions or curriculum at the individual level (if applicable)
   - Other uses that have been discussed such as using the assessment as part of a placement test?

   The validity of the assessment is always directly tied to its ability inform specific inferences and decisions. How do we define and protect the inference space (i.e., the range and kinds of inferences the assessment and the data will support) to avoid misuse of the assessment?

3. **Oral reading fluency**
   Like EGRA, OLA uses words-correct-per-minute (wcpm) to capture oral reading fluency.

   Is there a better way to document fluency or automaticity? Should a prosody measure be included?

4. **Setting cut-points and benchmarks**
   Does wcpm provide a reliable measure for setting basic cut-points and benchmarks for determining reading levels or grade equivalencies?

   OLA has three subtests with wcpm. What combination of these subtests is most desirable for setting benchmarks?

   Are there other sub-tasks or measures that could be used for determining cut-points and benchmarks that would have strong validity and reliability?

   What additional information would be useful for setting levels and defining the benchmarks within and between levels?

5. **Real-life reading items**
   Do the real-world items contribute anything to our understanding of the reading development of this population?

   Do these items supply meaningful information about literacy skills to host country governments and stakeholders? Project teams? Youth themselves?

   Should there be more real-life reading items for very beginning readers? For example, there's a drop-off in scores between the building signs and the bank hours. Should additional items be included?

   In some countries, such as Mali, there are no real-life reading items in the local environment in the language being tested (Bambara). What substitutions might be appropriate? What are the implications of changing this section of the assessment?

6. **Adaptation to new languages and/or country contexts**
   Is our adaptation process effective and efficient? (Reference: Memo 1 – Adapting OLA for languages, grade level equivalencies and contexts)

   Any recommendations?

7. **Alphabetic languages**
   Looking across our experience with Liberia, Rwanda, and Mali – can the set of sub-tasks (i.e., letter naming, word reading, etc.) work for most alphabetic languages?

   What are some considerations that we should keep in mind if we want to move into languages that use other alphabets?

8. **Data review and analysis**
   Are we on the right track with analysis and basic information from the data review regarding:
   - reliability of forms
   - IRR among test administrators
   - determining factor structure
   - equating forms
   - item analysis
   - examining potential differences in performance of key subgroups (e.g. gender, grade) using both descriptive statistics and differential item functioning
   - analyses to determine which learners appear to be profiting from the instruction and which are not
   - looking at attendance data
   - getting funding for small studies to analyze questions raised by OLA data

   What are your comments, suggestions and recommendations?

9. **Population distribution**
   Given that you've seen data only from the Liberia baseline, does OLA distribute this population according to the relevant early reading skills accurately and sufficiently? (Is it scalable?)

10. **Diagnostics**
    Does OLA provide enough diagnostic information to be helpful to program planners, teacher trainers, and curriculum developers?

    What is missing?

11. **Administrator Training and Test Administration Procedures**
    Are our processes sufficient for training test administrators and conducting the assessments?
    (Reference: Memo 4; administration guidelines)

    Any suggestions?

12. **General recommendations**
    What are your recommendations for improving design and content?

    What other information, if any, do you recommend that we collect through questionnaires or other data gathering to further understand and validate uses of OLA?

    Do you have recommendations for future piloting?

    Recommendations for improvements to the administration guidelines?

    Other recommendations?

13. **Next steps for the OLA**
    What are the priority next steps?

Appendix 2: Members of the External Review Team

Dr. Carolyn Adger is a Senior Fellow at the Center for Applied Linguistics. As an expert in sociolinguistics, she has worked on issues of language in society, including language use and language learning in educational settings. Her publications focus on the education of English language learners and speakers of vernacular dialects

Mr. Manuel Cardoso is a Program Specialist at the UNESCO Institute for Statistics (since 2005) and has worked at the Literacy Assessment and Monitoring Program (LAMP) since 2007, coordinating it since 2012, among other learning outcomes projects. With LAMP, Mr. Cardoso has worked on a large number of large-scale international adult literacy assessments in Africa, Latin America and Asia. In addition, he worked for eight years under Uruguay's National Administration of Public Education in projects including Uruguay's first-ever large-scale test in 1996 (a census of 6th graders), sample-based assessments for five grades, and the entry exam into teacher training. Mr. Cardoso has also taught at two universities in Uruguay and is the co-author of a number of publications.

Dr. Mary Beth Curtis is a professor at Lesley University and founding Director for Special Education. She is an expert in remedial reading and the impact of childhood trauma on learning. Dr. Curtis is the author of numerous articles on reading diagnosis and remediation, the role of vocabulary in comprehension, and the reading skills of at-risk teens and adults. She is a member of the Adult Literacy Research Working Group and was Lesley's Principal Investigator on a research project for improving the instruction of adult basic education intermediate readers, conducted in collaboration with Harvard University and Soliloquy Learning.

Dr. Jeff Davis is a trained psychometrician who has worked in the areas of research, management and policy development. Dr. Davis has a wide array of experience working domestically and internationally designing and leading major education assessments in literacy and math for formal and non-formal settings. He is an expert in quantitative methods and has advised the development of a number of large-scale assessments, including the Early Grade Reading Assessment and Early Grade Math Assessment. He has also advised the Global Partnership for Education in the area of assessment and policy.

Dr. Irwin Kirsch is the Director of the Center for Global Assessment at Education Testing Service (ETS). He has directed a number of large-scale assessments in the area of literacy including the National Adult Literacy Survey, the NAEP Young Adult Literacy Survey and the Adult Education Program Study with the U.S. Department of Education. He currently chairs the Reading Expert Group for the Programme for International Student Assessment (PISA), and oversees the development and implementation of the Programme for the International Assessment of Adult Competencies (PIAAC), a new computer-delivered assessment of adult competencies for the OECD.

Dr. Cristine Smith Crispin is a professor at the Center for International Education at the University of Massachusetts, Amherst and is an expert in literacy, adult and nonformal

education both domestically and internationally. Dr. Crispin has worked at the National Center for the Study of Adult Learning and Literacy (NCSALL), the U.S. Department of Education's only research center focused on adult basic education, as well as serving as the Director of World Education's literacy programs in South Asia. She is currently the Principal Investigator of the Adult Transitions Longitudinal Study (ATLAS), a $1 million, five-year social research project in New England funded by the Nellie Mae Education Foundation.

Dr. Nancy Clark-Chiarelli is a Principal Investigator/Research Scientist at EDC with over 35 years of experience dedicated specifically to language and literacy initiatives from pre-K through 12th grade. Dr. Clark Chiarelli has provided technical assistance to a large portfolio of domestic and international projects geared at improving teaching practices and teacher support systems while providing literacy expertise. She has been the lead developer on several projects for the National Public Broadcasting System where she designed the Reading and Language site for PBS Parents as well as professional development literacy courses for PBS's TeacherLine. Dr. Clark-Chiarelli is an author of the Early Language and Literacy Classroom Observation K-3, a widely-used literacy tool published by Brookes Publishing. She is a faculty member at Wheelock and Simmons Colleges, where she teaches courses in language, literacy, and special education.