REL SOUTHWEST Regional Educational Laboratory Southwest

Technical Memorandum: Early Learning Inventory (ELI) Study Analysis Methods

Date: October 20, 2022

To: Lauren Jenks-Jones, Oklahoma State Department of Education

From: Katie Dahlke, Principal Researcher; Yibing Li, Senior Researcher; Jingtong Pan, Senior Researcher; Maria Dracopoli, Research Associate; Alison Hauser, Researcher; Janice Keizer, Senior TA Consultant

REL Southwest

Subject: Early Learning Inventory (ELI) Study Analysis Methods

Introduction

With support from Regional Educational Laboratory (REL) Southwest, the Oklahoma State Department of Education (OSDE) developed and released an Early Learning Inventory (ELI) as an optional tool for Oklahoma public school kindergarten teachers during the 2021/22 school year. OSDE adapted the ELI from New Mexico's validated Early Childhood Observation Tool to align with Oklahoma Academic Standards. OSDE intends the ELI to serve as a resource for teachers to better understand their students' competencies and skills at the beginning of the school year, to individualize instructional activities on the basis of this information, and to track students' progress throughout the year (figure 1).

REL Southwest is now supporting OSDE with a study to examine teachers' implementation of the ELI and to obtain evidence of the ELI's validity in the Oklahoma context. The implementation study includes a sample of 44 kindergarten teachers across 12 districts who used the ELI during the 2021/22 school year. OSDE shared the opportunity with all district administrators through an e-newsletter announcement and followed up with district contacts to share more information about the ELI. Twelve districts and 44 teachers requested to participate in the study. Therefore, all interested districts and teachers were included. The 12 districts included both large and small districts from different areas of the state; however, the teachers in the study were not a representative sample of Oklahoma kindergarten teachers. For example, the ELI study sample has more teachers with master's degrees compared to teachers in the state (25.0% compared to 16.6%) and fewer teachers with a non-traditional certification (13.5% compared to 18.3%). The validation study includes student-level ELI data from 851 students in these teachers' classrooms as well as two additional "general ELI users" who are kindergarten teachers who completed the ELI training but were not part of the implementation study.

Figure 1. ELI research and development process

This memo provides a summary of the research questions, data sources, and analysis methods for this study.

Research questions

The ELI study was designed to address 10 research questions:

Validation research questions

1. What domains of students' learning and development does the ELI validly measure?
2. Do any of the ELI indicators exhibit potential bias for students with different characteristics?
3. Do teachers use rating categories for each ELI item as intended?
4. To what extent does the ELI provide information about individual student abilities?

Implementation research questions

5. To what extent do the ELI training and resources prepare kindergarten teachers to use the ELI?
6. How do kindergarten teachers report administering the ELI in their classroom, and do they report using the ELI data to inform instruction?
7. What are kindergarten teachers' perceptions about the feasibility and value of using the ELI in their classroom?
8. What are the key facilitators and challenges for kindergarten teachers implementing the ELI with fidelity?
9. What improvements could be made to the ELI training, measure, and technology platform to increase feasibility and fidelity?
10. What motivated districts to use the ELI?
    a. *What are administrators' perceptions about the value of the ELI?*

This memo describes the analyses REL Southwest conducted to address 9 of the 10 research questions (1–4 and 5–9) using data collected in fall 2021. REL Southwest will subsequently conduct analyses to address research question 10 with data collected in the spring of 2022 and this will be described in a separate technical memo.

Data sources

REL Southwest used data from five main sources to address the research questions: student-level ELI data, student-level Reading Sufficiency Act (RSA) assessment data, student-level enrollment and student characteristics administrative data, teacher surveys, and teacher focus groups.

REL Southwest used ELI assessment data, RSA assessment data, and student enrollment and characteristics from the OSDE to address the validation research questions (research questions 1–4). These data sources are described in more detail in this section.

Student-level Early Learning Inventory assessment data

REL Southwest obtained extant, deidentified student-level ELI indicator ratings from fall 2021 to conduct the analysis for four of the validation research questions (research questions 1–4). The ELI includes 26 indicators each with six rating categories that follow a learning progression. OSDE provided a pseudo-identification (ID) number for students and their data file linked the student data to classroom teachers. By linking student data to classroom data, the analyses can appropriately account for the nesting of students within teachers.

Reading Sufficiency Act assessments

REL Southwest obtained extant, deidentified beginning-of-year, student-level RSA assessment data for a subsample of the students for whom there were ELI data. Oklahoma has seven different RSA assessments that districts may choose to administer. OSDE was only able to obtain student-level data from the Renaissance Star Early Literacy, which is the most commonly used RSA assessment across the state. The RSA assessments measure students' reading skills (including phonemic awareness, phonics, reading fluency, vocabulary, and comprehension) in kindergarten. REL Southwest used the RSA total scaled scores (summarizing student performance across all items and standardized) to measure the concurrent validity of literacy-related domains in the ELI (research questions 1 and 4). OSDE provided pseudo-IDs for students linked to the RSA data to allow for merging with the ELI data.

Student enrollment and characteristics from the Oklahoma Department of Education data

REL Southwest obtained de-identified data about kindergarten students in classrooms using the ELI in fall 2021 including race/ethnicity, English learner student status, free or reduced-price lunch eligibility, and special education status. REL Southwest used these data in two primary ways. First, the data were used to summarize the sample descriptively and assess the extent to which our sample is generally representative of Oklahoma kindergarten students in terms of observable characteristics. Second, REL Southwest used the data as part of the validation analysis to explore variation in response patterns by teachers on the ELI by student demographics (specifically, differential step functioning; research question 2). OSDE provided pseudo-IDs for students to allow for merging with the ELI and RSA data.

Teacher surveys

REL Southwest used teacher surveys and teacher focus groups to address the implementation research questions (research questions 5–9). These data sources are described in more detail in this section.

With support from REL Southwest, OSDE administered three teacher surveys as part of the ELI study to date. The first survey was a brief baseline (pre-training) survey administered at the time of recruitment using an online survey platform (in spring 2021) preceding implementation of the ELI. The goal of the survey was to collect baseline data on teacher self-efficacy related to formative assessment administration and the use of data to inform instruction. The second survey was a brief post-training survey administered at the end of the ELI training (in summer 2021). The goal of the post-training survey was to identify participants' satisfaction with the training and to determine what training topics and materials could be improved. The third survey, also using an online survey platform, was administered after the beginning-of-year ELI administration (in November 2021). The third survey includes the same items as the baseline pre-training survey to examine change over time as well as several items related to implementation considerations (for example, time to log ratings in the ELI system). The teacher survey data were used to

help answer implementation (and feasibility) questions and to help inform future efforts to scale up the use of the ELI across the state.

Teacher focus groups

REL Southwest conducted two teacher focus groups with six kindergarten teachers in total who participated in the ELI study in November 2021. OSDE also conducted six focus groups with the remaining 38 teachers. The data from these focus groups were in the form of transcribed audio recordings. The topics of discussion in the focus groups align with the first five implementation research questions. REL Southwest did not link an individual teacher's focus group data to their teacher survey data because of the small number of focus group participants.

Analysis methods

This section describes the analytic approaches REL Southwest used to address the research question. In addition, this section describes data preparation steps.

Student-level data preparation

The research team merged student-level and teacher-level data files received by OSDE. For the student-level data file with student characteristics, teacher-provided ratings on the ELI, and the RSA scores, REL Southwest ran descriptive analyses on all the variables, means for continuous variables, frequency distributions for categorical variables, minimum values, maximum values, and percentage missing, to confirm that the attributes of the information are within reason. In addition, the research team reviewed the file to remove any duplicate observations, observations with invalid ratings, and observations with birthdates outside of the range of kindergarten ages.

Teacher survey data preparation

Nonresponse analyses and weighting

REL Southwest examined the extent to which there is nonresponse. Nonresponse can occur at the unit and item levels. Unit nonresponse occurs when a teacher does not complete the survey at all. Item nonresponse for the teacher-level data occurs when a teacher completes part of the survey but skips some items, or for the student-level data if some information is provided for ELI indicators but not all.

Unit nonresponse

Study participants agreed to complete the ELI and the teacher surveys when they signed up for the study. There was no unit nonresponse for the teacher surveys at pre-training, post-training, and the first follow-up.

Item nonresponse

REL Southwest examined item-level nonresponse to determine whether any items have less than an 85 percent response from respondents. Item-level nonresponse was calculated as the number of cases providing a valid response to an item divided by the number of cases eligible for that item. If items had a nonresponse rate of 15 percent or more, REL Southwest determined whether data from these items should be included in the analyses (explicitly noting the response rate so that findings can be interpreted within that context) or excluded from the analyses. This determination was informed by looking at the pattern of item-level nonresponse and looking for potential bias by examining the distribution of characteristics known for the entire sample, by respondents versus nonrespondents on the particular item. There were four items that had item nonresponse of 15 percent or more. These items were not excluded from analyses, but they were flagged as items having low response rates in the findings slides and the

characteristics of respondents and non-respondents were presented in the supplemental slides. The respondents and non-respondents were similar on most of the characteristics, with some minor differences, such as years of experience. A greater proportion of the non-respondents had three or more years of experience compared to responders.

Construct validity analyses

REL Southwest conducted psychometric analyses to address research questions 1–4 that involve statistical analysis of the student-level data, including the data on student characteristics, the ELI data, and RSA assessment data. These methods followed those used in the published REL Southwest study that examined the validity of the New Mexico Kindergarten Observation Tool (Dahlke et al., 2017).

Analyses for Research question 1. What domains of students' learning and development does the ELI validly measure?

REL Southwest conducted analyses to examine the internal factor structure, internal consistency, and convergent validity of the ELI. Each item in the ELI was developed to measure a skill that belongs within one of six content domains: Physical Development, Health, and Well-Being; Literacy; Numeracy; Scientific Conceptual Understanding; Self, Family, and Community; and Approaches to Learning. In theory, therefore, the ELI instrument may provide scaled scores representing each of the six domains. However, prior analyses of the New Mexico Kindergarten Observation Tool found the indicators fit best within just two domains. An empirical examination of the latent constructs in the instrument determined whether the instrument separately measures the six domains as intended, or whether some indicators group differently into combined domains or do not contribute reliably to any domain score. Furthermore, this empirical examination indicates whether the ELI produces a total score that is a reliable measure of children's overall knowledge and skills across domains. Additionally, REL Southwest examined the concurrent validity of the ELI with the RSA.

Internal factor structure

REL Southwest examined the internal factor structure of the ELI to uncover the number of distinct domains or groups of ELI indicators that best describe the data. To conduct this empirical examination of the latent domains measured by the ELI, REL Southwest conducted both exploratory and confirmatory factor analysis with the ELI data using *Mplus* 6.0. Confirmatory factor analyses can provide information on the degree to which the relationships among indicators and domains align with the intended domains proposed for scoring and use. Exploratory factor analyses can help identify alternative domains for scoring and use in which the indicators are more correlated with one another.

Because all the ELI indicators are on a six-point ordinal scale, REL Southwest conducted factor analysis based on the weighted asymptotic covariance matrix of the polychronic correlation using a robust weighted least square estimator (WLSMV; Muthén & Muthén, 2010). Missing data were accounted for with the WLSMV estimator, which treats missingness as a function of the observed covariates. REL Southwest considered a group of fit indices to determine model fit: χ^2 goodness-of-fit statistic, comparative fit index (CFI), the Tucker–Lewis index (TLI), the root mean square error of approximation (RMSEA), and weighted root mean residual (WRMR). Nonsignificant values of χ^2, RMSEA value smaller than .08, and CFI values greater than .90 would indicate that a hypothesized model fits the data (Hu & Bentler, 1999). As χ^2 is sensitive to sample size, we drew on the other fit indices to decide model fit. Multilevel models were estimated using a saturation approach[1] to account for students nested in teachers. When comparing nested models with increased equal constraints across multiple groups, we

[1] A model with a perfect fit to the data (i.e., a saturated model) at the between level and looking at within-level structures.

applied the DIFFTEST command in *Mplus* 6.0 on all chi-square difference tests. In addition, changes in RMSEA, CFI, and TLI greater than .01 were considered significant as discussed by Chen (2007).

The full sample of 853 students was divided into two random subsamples of comparable sizes, 425 students (sample 1) and 426 students (sample 2). First, REL Southwest conducted an exploratory factor analysis (EFA) using the first random subsample to determine the optimal factor structure. Then, REL Southwest conducted confirmatory factor analyses (CFAs) of the developer's proposed factor structure and of the structures identified through EFA to test the model fit in a different sample of students using the second random subsample. The CFA, using a separate subsample of the data, served as a way to validate these empirical latent constructs (Thompson, 2004).

For the EFA with the first random subsample, squared multiple correlations were used as the initial communality estimates. We applied oblimin rotation to obtain the factor structure. We retained items with factor loadings above a threshold value of 0.4 (Stevens, 2009) and removed items with cross-loadings (that is, items load on multiple constructs with factor loadings above 0.4). REL Southwest picked the final factor structure based on six criteria: the extent to which the model fit indices were within the acceptable threshold, the hyper-plane count (Yates, 1987), the total number of nonsalient loadings based on a threshold value of 0.4 (Stevens, 2009), the number of cross-loadings (smaller number suggests better fit), the closeness to a simple structure (Fabrigar et al., 1999) (the simpler the structure, the better), the internal consistency of each of the factors, and the meaningfulness of each factor (Gadermann, Guhn, & Zumbo, 2012). After determining the optimal factor structure from the EFA in the first random subsample (by selecting the model with the best fit to the data), REL Southwest performed CFA with the second random subsample. In addition, REL Southwest ran two additional models to compare model fit statistics with the optimal factor structure derived from the EFA, including a CFA model based on including the six-factor structure proposed by the developers, and a CFA model for a single latent construct of school readiness. Differences in RMSEA, CFI, and TLI greater than .01 were considered significant as discussed by Chen (2007).

Internal consistency

To determine the reliability of the ELI, the study team examined the internal consistency for each of the factors obtained from the factor analyses. Internal consistency describes the magnitude of the relationship among indicators within a domain.

REL Southwest examined the polychoric correlation matrix for indicators associated with the same factor but did not find any indicators that have low correlations with others. REL Southwest also examined the item-total correlations for each item. Next, the REL Southwest team ran internal consistency analyses to assess the unidimensionality for each group factor based on McDonald's Omega (McDonald, 1999). To assess whether the internal consistency of the latent constructs in the ELI would be improved by removing any indicators, REL Southwest calculated the internal consistency for each latent construct after removing each item but did not find any indicators that would lead to a higher internal consistency upon removal. REL Southwest estimated the Rasch item reliability and Rasch person reliability, indicators of whether the data allow grouping item (Rasch item reliability) or people (Rasch person reliability) into multiple categories. Rasch person reliability is largely driven by the targeting/alignment between the item difficulties and the target population ability distribution. By contrast, alpha coefficients are largely driven by the extent to which the individual item responses (not measures) correlate with each other, regardless of the item-person targeting.

Convergent validity analyses

Convergent validity analysis can provide additional evidence of whether an assessment measures the intended construct by looking at the strength of the correlation with other independent, reliable, and valid

assessments. Convergent validity evidence is when two assessments of the same or similar construct have a strong (or stronger) relationship.

To examine convergent validity, the study team examined correlations between RSA assessment scores and scores on ELI indicators that were designed to measure the same underlying construct, that is, early literacy skills among the 134 students who had both RSA data and ELI data. Due to the relatively small subsample size, findings from the convergent validity analysis may not accurately reflect the full sample and should be interpreted with caution.

Examining correlations between the ELI domain with literacy indicators and the RSA assessment scores would provide information about the degree of convergence of the ELI domain with literacy indicators with the RSA measures, often referred to as "convergent validity." Spearman's rank-order correlations were used to examine these item-level relationships because the item-level ELI scores were ordinal and may not be normally distributed, and Full Information Maximum Likelihood (FIML) estimation was used to account for missing data. For domain-level correlations, REL Southwest used Pearson correlations. We used correlations of 0.70 and higher to indicate evidence of good convergence—that is, the two scales measured a highly similar construct (Di Iorio, 2005). Null hypothesis significance testing was used to test whether each ELI literacy item and the RSA assessment composite score was independent.[2]

Analyses for Research question 2. Do any of the ELI indicators exhibit potential bias for student subgroups?

Indicator bias occurs when groups of students with different characteristics but with equal ability differ in the likelihood of having an indicator positively endorsed by the teacher. In other words, teachers' assessments of students' abilities (their ratings) are not based exclusively on the child's ability for these subgroups.

To evaluate if the ELI is valid and unbiased across subgroups, REL Southwest first conducted multiple-group confirmatory factor analysis to determine whether we can achieve factor invariance across subgroups (that is, by gender, English learner status, eligibility for free or reduced-price meals, special education status, and by race/ethnicity). Configural invariance and scalar invariance were examined in sequence. A stronger case of invariance is established when there was no sizable difference considering the model fit statistics (Muthén & Asparouhov, 2002), indicated by either the nonsignificance of the chi-square difference test (Marsh & Grayson, 1994) or the minimal shift in CFI and RMSEA (change of CFI less than 0.01 and change of RMSEA less than 0.015; Cheung & Rensvold, 2002). Note that polychoric correlation matrices require that data contain all values of each categorical variable included in the analysis. Due to the relatively small sample sizes for the English learner student group, special education student group, and racial minority student groups, not all indicators have all values of each categorical variable for the smaller student groups. As a result, the multiple-group confirmatory factor analyses across these student groups did not converge and therefore results could not be obtained for these characteristic contrasts. We do, however, have evidence for factor invariance for gender and for free or reduced-price lunch status.

Next, the research team examined differential item functioning (DIF). Different from the CFI approach which looks at all items together as a single construct and examines group bias by construct, DIF is a way to assess measurement bias at the item level, which occurs if subgroups have the same latent construct level but have differing patterns of response to a specific item measuring that latent construct. For example, an item would exhibit measurement bias for Hispanic children if the teachers of Hispanic children tend to rate the Hispanic children substantially lower on that specific item than non-Hispanic children of otherwise equivalent kindergarten readiness. In addition, ELI ratings may be different in fall

[2] Bonferroni corrections for multiple hypothesis testing will be used to reduce the likelihood of a type 1 error.

2021 than in other years due to students' unique early childhood experiences during the COVID-19 pandemic. Evidence of differential item functioning raises concerns about the validity of the instrument for children in the affected subgroup (Greenberg, Penfield, & Greenfield, 2015; Qi & Marley, 2009).

To determine whether any of the indicators exhibit bias for teachers' ratings of subpopulations of students, differential step functioning (DSF) analyses with Rasch modelling was used. DSF calculates whether the probability of attaining each rating level on each item is statistically equivalent across subgroups after adjusting for their kindergarten readiness based on all other indicators (Wright & Stone, 1979). Compared with traditional DIF analyses, DSF can help pinpoint precisely which score level is responsible for an observed DIF effect and the potential causes of the DIF effect (Penfield, 2007). To be more specific, REL Southwest used the log odds-ratio statistics to screen DSF, which did not require an adequate model fit assumption compared with item response theory-based methods (Cohen, Kim, & Baker, 1993; Liu & Agresti, 1996; Penfield & Algina, 2003). An absolute value of 0.64 or more for the log odds-ratio statistics would indicate potential bias for an item (Penfield, Alvarez, & Lee, 2009; Zieky, 1993). These analyses used complete cases only.

Evidence of DSF does not directly translate into bias of the item (Walker, 2011), for example, when differing patterns of response to a specific item are results of true differences among student groups. It is common in DSF studies to identify some indicators that suggest DSF. To determine whether the presence of DSF indicates item influence (when teacher ratings of children from different groups have differing probabilities of endorsing an item due to true differences in the underlying ability) or item bias (when teacher ratings of children from different groups have differing probabilities of endorsing an item due to some characteristic of the test item that is not relevant to the test purpose), subject matter experts should be consulted (Camilli & Shepard, 1994; Sandoval & Miille, 1980). REL Southwest flagged indicators based on DSF analyses and will suggest that OSDE may wish to conduct further review and consultation to determine whether true bias exists or if the differences reflect a true difference for students with those characteristics.

Analyses for Research question 3. Do teachers use rating categories for each ELI item as intended?

REL Southwest examined the performance of each item using Rasch analyses. Rasch analyses can provide information on whether teachers use higher-level indicator ratings to characterize students with the higher overall ability levels and lower-level indicator ratings to characterize students with lower overall ability levels. We proposed to use Rasch models because they are simpler and more intuitive conceptually and can allow specification of non-linear relationships between observed and latent variables.

For each dimension in the ELI, REL Southwest fitted a partial credit polytomous Rasch model to the indicators. The Rasch model estimates the threshold parameters corresponding to the transition in the latent trait levels that are represented by two adjacent skill points (van der Linden & Hambleton, 2013). REL Southwest used joint maximum likelihood estimation to account for missing data. Items were flagged as problematic if, for example, the threshold parameter for the jump from a rating of 1 to a rating of 2 was less than the threshold for the jump from a rating of 2 to a rating of 3 (Andrich, 2010; Bond & Fox, 2007).

Inappropriate skill ordering, multidimensional responses, and skewed response patterns all can contribute to threshold reversals (Adams, Wu, & Wilson, 2012; Andrich, de Jong, & Sheridan, 1997). Adams and colleagues (2012) and Bond and Fox (2007) suggest that reversals indicate that an item is malfunctioning and thus should be revised to ensure correct ordering of the rating categories.

Analyses for Research question 4. To what extent does the ELI provide information about individual student abilities?

REL Southwest conducted an examination of teacher's variance in ELI indicator ratings to provide information about the amount of clustering in ELI ratings within classrooms. An examination of assessor variance provides reliability and validity evidence of the use of a given measure (Waterman, McDermott, Fantuzzo, & Gadsden, 2011). Assessor variance (with the ELI, the assessor is the teacher) is the amount of variation in ratings that is attributable to the assessor or other factors and not to the children's abilities or skills. Although ideally the score should solely reflect a student's performance, when an assessment is administered by an assessor, some of the information captured may reflect the assessor's perspective or bias or lack of training using the assessment.

The research team applied multilevel modeling to estimate the proportions of score variability that were attributable to children and assessors. A two-level unconditional model (student, teacher-classroom) was fit separately for each indicator and validated factor, and REL Southwest calculated the variance partition coefficient (Goldstein, Brown, & Rasbach, 2002) or the proportion of the variance explained by the teacher and classroom level of the model. The variance partition coefficient (VPC) can be calculated in the general form below:

$$VPC = \sigma_u^2/(\sigma_e^2 + \sigma_u^2)$$

where σ_u^2 is the between teachers/classrooms variance and σ_e^2 is the residual variation between children.

We acknowledge that the between classroom variance could be explained by a number of factors. Rater error is one explanation; however, another potential explanation is that the variance is caused by sorting of children with similar ability into the same classrooms. To investigate this potential explanation of children clustered by similarities, REL Southwest used the same procedure as described above to partition the between and within classroom variation of an independent direct assessment (specifically, the RSA assessment because it is a standardized test with established validity). Direct assessments have fewer opportunities to introduce rater error. As such, the amount of between classroom variance from the RSA assessment would signal the amount of variation in children's abilities between classrooms. Finally, the difference in the variance partition coefficient between the RSA assessment and the ELI assessment would indicate the between classroom variance introduced by rater error.

Implementation analyses (analyses for research questions 5-9)

Implementation analyses were conducted with the teacher-level data. The research team used two data sources to address each of the five implementation research questions (research questions 5–9): teacher surveys and teacher focus groups. For each research question, the team looked at response distributions for the relevant survey items and conducted basic descriptive summaries of survey responses (see table 1). For example, we conducted cross-tabulations to compare responses between early career and experienced teachers. For the self-efficacy items, the research team examined change over time. We conducted thematic analysis of the focus group data. The instruction for coders can be found in appendix D.

5. To what extent do the ELI training and resources prepare teachers to use the ELI?
6. How do teachers report administering the ELI in their classroom, and do they report using the ELI data to inform instruction?
7. What are teachers' perceptions about the feasibility and value of using the ELI in their classroom?
8. What are the key facilitators and challenges for teachers implementing the ELI with fidelity?

9. What improvements could be made to the ELI training, measure, and technology platform to increase feasibility and fidelity?

Survey analyses

Individual items

REL Southwest tabulated responses to each survey item and developed bar graphs to show the distribution of responses overall. The research team examined distributions separately for new and experienced teachers (that is, teachers with three or more years of experience). Table 1 outlines the alignment between the survey item sets and research questions. The three surveys are included in this document in appendices A, B, and C.

Table 1. Survey item sets and research question crosswalk

Research questions	Pre-training survey	Post-training survey	Post-ELI survey
5. To what extent do the ELI training and resources prepare teachers to use the ELI?	#1	#1	#1, 2
6. How do teachers report administering the ELI in their classroom, and do they report using the ELI data to inform instruction?	n/a	n/a	#3–8
7. What are teachers' perceptions about the feasibility and value of using the ELI in their classroom?	n/a	n/a	#9–11
8. What are the key facilitators and challenges for teachers implementing the ELI with fidelity?	n/a	n/a	#12
9. What improvements could be made to the ELI training, measure, and technology platform to increase feasibility and fidelity?	n/a	#2–4	#13, 14
Demographic characteristics	#2–8	n/a	n/a

n/a is not applicable.

Teacher self-efficacy to use formative assessment

To examine changes in teacher self-efficacy to use formative assessment practices over time (implementation research question 5), REL Southwest created a composite measure as the average of the relevant items (items 1a–1n on the pre-training and post-ELI survey). Next, we used a paired samples *t*-test to examine if changes over time are statistically significant:

$$t = \frac{x_2 - x_1}{\sqrt{(SE_2{}^2 + SE_1{}^2)}}$$

where x_1 and x_2 are the estimates being compared and SE_1 and SE_2 are the corresponding standard errors.

Open-ended items

For the open-ended survey items, the research team examined responses for potential coding into a category of findings for presentation in the report. Two researchers served as coders, using Excel to support their coding. To ensure inter-rater reliability, all surveys were double coded. The team reviewed

codes by both coders and found that there was adequate level of agreement through observation of at least 75 percent agreement.

Teacher focus group analyses

REL Southwest obtained transcripts of the focus groups and conducted an analysis of the transcripts. REL Southwest used a spreadsheet to code focus groups transcripts. The coding of the data involved reading and evaluating each comment in the transcript and labeling sections of the transcripts with apriori and emergent theme codes. *A priori* codes included implementation research questions 5–9 at the parent code level as well as subthemes at the child code level. Examples of the *a priori* subthemes include codes about ELI ease of use; data dashboard likes and dislikes; types of support for data use and training resources that were useful or not useful. A second coder from the team reviewed each transcript to ensure that all comments were coded and to determine if any codes should be reconsidered. After all data were coded and reviewed, REL Southwest summarized the main themes into summary tables that include exemplary quotes as necessary to enrich understanding.

Limitations of the study design, data, or analyses

The limitations of the study with respect to the teacher survey include considerations of the reliability of self-reported data (such as the risk of bias in reporting information to the state education agency) and potential considerations that the ELI study sample is purposive and not random, and therefore may not be exactly representative of the distribution of all teachers. For example, the ELI study sample has more teachers with master's degrees compared to teachers in the state (25.0% compared to 16.6%) and fewer teachers with a non-traditional certification (13.5% compared to 18.3%). Self-reported data may not present a full, accurate picture because the data reflect teachers' perceptions, not independent measures. In addition, when teachers report information to the state education agency, they may intentionally or unintentionally over-report the extent to which they are using a tool such as the ELI.

A limitation of the student data with respect to the ELI data is that the validation analyses were conducted with students' data within a small sample of classrooms, schools, and districts and during the first year of implementation, as such the results may not generalize to all districts across the state nor to future administrations of the measure when any potential preliminary implementation challenges are addressed.[3] In addition, ELI ratings may be different in fall 2021 than in other years due to students' unique early childhood experiences during the COVID-19 pandemic. For example, there could be differential effects of the pandemic on learning and learning loss among subgroups of students that, in turn, might lead to systematic differences/non-invariance of the ELI. Future validation studies will be needed with a sample that is representative of students across the state and after any changes are made to the ELI measure and/or training based on the study.

References

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, 72*, 547–573. http://eric.ed.gov/?id=EJ982111

Andrich, D. (2010). Understanding the response structure and process in the polytomous Rasch model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 123–152). New York, NY: Taylor & Francis.

[3] The state's 2021/22 kindergarten race/ethnicity data was reported such that children who were identified as Hispanic did not have a race designation. Therefore, we were not able to make a clear comparison between the state data and the ELI study data which indicated student race and ethnicity separately.

Andrich, D., de Jong, J. H., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 59–70). Munster, Germany: Waxmann.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling, 14, 464-504. doi:10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. http://eric.ed.gov/?id=EJ651463

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*, 335–350. http://eric.ed.gov/?id=EJ485766

Dahlke, K., Yang, R., Martínez, C., Chavez, S., Martin, A., Hawkinson, L., Shields, J., Garland, M., & Carle, J. (2017). *Scientific evidence for the validity of the New Mexico Kindergarten Observation Tool* (REL 2018–281). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://ies.ed.gov/ncee/edlabs.

Di Iorio, C. K. (2005). *Measurement in health behavior: Methods for research and evaluation*. San Francisco, CA: Jossey-Bass.

Duncan, G. J., & Magnuson, K. A. (2005). Can family socioeconomic resources account for racial and ethnic test score gaps? The Future of Children, 15(1), 35–54. https://eric. ed.gov/?id=EJ795842

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation, 17*(3), 1–13. http://eric.ed.gov/?id=EJ977577

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics, 1*, 223–231.

Greenberg, A. C., Penfield, R. D., & Greenfield, D. B. (2015). *Differential item functioning in a large-scale preschool science assessment*. Manuscript in preparation.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. http://eric.ed.gov/?id=EJ576570

Liu, I.-M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*, 1223–1234.

Marsh, H. W., & Grayson, D. (1994). Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. *Structural Equation Modeling, 1*(2), 116–145.

McDonald, R. P. (1999). Test theory: A unified treatment. Lawrence Erlbaum.

Muthén, B., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Retrieved January 27, 2016, from http://www.statmodel.com/download/webnotes/CatMGLong.pdf.

Muthén, L. K., & Muthén, B. O. (2010). Growth modeling with latent variable using Mplus: Advanced growth models, survival analysis and missing data. *Mplus Short Courses*.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 44*(3), 187–210. http://eric.ed.gov/?id=EJ772970

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*, 353–370.

Penfield, R. D., Alvarez, K., & Lee, O. (2009). Using a taxonomy of differential step functioning form to improve the interpretation of DIF in polytomous items. *Applied Measurement in Education, 22*(1), 61–78. http://eric.ed.gov/?id=EJ822892

Qi, C. H., & Marley, S. C. (2009). Differential item functioning analysis of the "Preschool Language Scale–4" between English-speaking Hispanic and European American children from low-income families. *Topics in Early Childhood Special Education, 29*(3), 171–180. http://eric.ed.gov/?id=EJ859162

Sandoval, J., & Miille, M. P. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology, 48*, 249–253. http://eric.ed.gov/?id=EJ225027

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York, NY: Routledge.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (2013). *Handbook of modern item response theory*. New York: NY: Springer Science & Business Media.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment, 69*(4), 364–376. http://eric.ed.gov/?id=EJ935633

Waterman, C., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2012). The matter of assessor variance in early childhood education—Or whose score is it anyway? *Early Childhood Research Quarterly, 27*, 46–54. http://eric.ed.gov/?id=EJ947506

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Press.

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. Albany, NY: State University of New York Press.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Lawrence Erlbaum.

Appendix A: Pre-Training Survey

Pre-training Survey: items administered before ELI training

1. **Please select your level of agreement for each of the following statements.** Select one answer for each row.

	Completely disagree	Disagree	Agree	Completely agree
a. I am effective at administering formative assessments.	○	○	○	○
b. I can effectively incorporate use what I learn about students through formative assessment into my instruction	○	○	○	○
c. Formative assessments are a useful tool to improve my practice.	○	○	○	○
d. In general, assessments help teachers plan instruction.	○	○	○	○
e. In general, assessments offer information about students that was already known.	○	○	○	○
f. In general, assessments help teachers know what concepts students are learning.	○	○	○	○
g. In general, assessments help teachers know what concepts students need to learn.	○	○	○	○
h. In general, assessments help teachers identify learning goals for their students.	○	○	○	○
i. Students benefit when teacher instruction is informed by assessment data.	○	○	○	○
j. I think it is important to use assessment data to make instructional decisions.	○	○	○	○
k. I like to use assessment data.	○	○	○	○
l. I find assessment data useful.	○	○	○	○
m. Using assessment data helps me be a better teacher.	○	○	○	○

2. **How many years of paid teaching experience, including this year, do you have as a certified teacher?**

_____

3. **How many years of paid teaching experience, including this year, do you have as a certified teacher *teaching kindergarten students*? Do not include years that you taught another grade level.**

_____

4. **What current state certification(s)/qualifications(s) do you hold?** Check all that apply.

- ☐ Traditional Certification (completed a state-approved teacher education program)
- ☐ Emergency Certification (temporary certification when there is no teaching license at the grade level or subject area)
- ☐ Alternative Certification (minimum of a bachelor's degree in any area but did not complete a state-approved teacher education program)
- ☐ Other (please specify):_____

5. **What is the highest level of education you have completed?**

- ☐ Bachelor's degree
- ☐ Some graduate coursework (no master's degree)
- ☐ Master's degree
- ☐ Doctorate or professional degree
- ☐ Other (please specify):_____

6. **I identify as:** (Select all that apply.)

- ☐ White/Caucasian Black/African American Hispanic/Latinx
- ☐ Black/African American
- ☐ Hispanic/Latinx
- ☐ Asian
- ☐ Native American
- ☐ Native Hawaiian/Pacific Islander
- ☐ Other (please specify):_____
- ☐ I prefer not to answer

7. Do you teach full-day or part-day kindergarten?

- ☐ Full-day kindergarten
- ☐ Part-day kindergarten

8. How many kindergarten students are currently enrolled in your kindergarten class(es)?

_____

9. How much planning time are you allotted in a normal school week this fall?

ENTER HOURS AND MINUTES: _____

10. As far as you know right now, how will your classroom be meeting when you start the school year this fall?

- o All virtual instruction
- o All in-person instruction
- o A hybrid model of in-person and virtual instruction

Appendix B: Post-Training Survey

Early Learning Inventory (ELI) Teacher Surveys

Post-training Survey: items administered at end of ELI training

1. Please select your level of agreement for each of the following statements. Select one answer for each row.

	Completely disagree	Disagree	Agree	Completely agree
a. I feel prepared to assess students on the 26 ELI indicators	◯	◯	◯	◯
b. I understand the ELI rubric categories.	◯	◯	◯	◯
c. I think the ELI rubric categories were easy to distinguish from one another.	◯	◯	◯	◯
d. I feel ready to implement the ELI in my classroom.	◯	◯	◯	◯
e. I feel prepared to use ELI data to inform instructional decisions for my students.	◯	◯	◯	◯

2. **Please indicate the extent to which you are satisfied with the ELI training.**

	Completely unsatisfied	Unsatisfied	Satisfied	Completely satisfied
a. Length of the training	◯	◯	◯	◯
b. Instructional style of the training	◯	◯	◯	◯
c. Materials used in the training (handouts)	◯	◯	◯	◯

3. What recommendations do you have, if any, to improve the *ELI training*?

_____

_____

_____

4. What recommendations do you have, if any, to improve the *ELI training materials*?

_____

_____

_____

Appendix C: Post ELI Survey

Post-pilot Survey: items administered after ELI pilot (first follow-up)

1. Please select your level of agreement for each of the following statements. Select one answer for each row.

	Completely disagree	Disagree	Agree	Completely agree
a. I am effective at administering formative assessments.	○	○	○	○
b. I can effectively incorporate use what I learn about students through formative assessment into my instruction	○	○	○	○
c. Formative assessments are a useful tool to improve my practice.	○	○	○	○
d. Assessments help teachers plan instruction.	○	○	○	○
e. Assessments offer information about students that was already known.	○	○	○	○
f. Assessments help teachers know what concepts students are learning.	○	○	○	○
g. Assessments help teachers identify learning goals for their students.	○	○	○	○
h. Students benefit when teacher instruction is informed by assessment data.	○	○	○	○
i. I think it is important to use assessment data to inform education practice.	○	○	○	○
j. I like to use assessment data.	○	○	○	○
k. I find assessment data useful.	○	○	○	○
l. Using assessment data helps me be a better teacher.	○	○	○	○

2. In your opinion, how important is the ELI for the following purposes?
For those that you think are not appropriate, please select "Not an appropriate use".

In your opinion, how important is the ELI for…	Not an appropriate use	Minimally important	Moderately important	Very important
a. establishing a baseline for student progress.	○	○	○	○
b. measuring growth over time.	○	○	○	○
c. recording data for the district.	○	○	○	○
d. determining what I will need to teach over the school year.	○	○	○	○
e. telling parents how their child is doing.	○	○	○	○
f. individualizing instruction.	○	○	○	○

In your opinion, how important is the ELI for…	Not an appropriate use	Minimally important	Moderately important	Very important
g. understanding students' strengths and needs	◯	◯	◯	◯
h. informing instructional activities	◯	◯	◯	◯
i. determining if a child needs an IEP	◯	◯	◯	◯
j. evaluating a teacher	◯	◯	◯	◯
k. evaluating a school	◯	◯	◯	◯
l. grouping students for instructional activities	◯	◯	◯	◯
m. Other (please specify): _____	◯	◯	◯	◯

3. For each of the following classroom grouping descriptions, indicate how frequently you gathered evidence of students' knowledge, behaviors, and skills to inform ELI indicator ratings. Select one answer on each row.

	Never/not often	Sometimes	Often
a. During **regularly-planned** whole group instructional time.	◯	◯	◯
b. During **regularly-planned** small group instructional time.	◯	◯	◯
c. During **regularly-planned** one-on-one instructional time.	◯	◯	◯
d. During whole-group instructional time **designed intentionally for the purpose of completing the ELI.**	◯	◯	◯
e. During small-group instructional time **designed intentionally for the purpose of completing the ELI.**	◯	◯	◯
f. During one-on-one instructional time **designed intentionally for the purpose of completing the ELI.**	◯	◯	◯

4. Please select the response that most closely represents how you used the ELI across the students in your classroom. Select one answer on each row.

	For no students	For a few students	For a majority of students	For all or nearly all students
a. I gathered evidence to inform ELI indicator ratings over the first 30 instructional days of the school year.	◯	◯	◯	◯
b. I generated student-level ELI reports.	◯	◯	◯	◯
c. I generated parent/student ELI reports.	◯	◯	◯	◯

5. Thinking back over the last four weeks, please indicate if or how often you completed the following activities. Select one answer on each row.

	Never	At least once	At least once a week	Daily or almost daily
a. I generated classroom- level ELI reports.	◯	◯	◯	◯
b. I met with an administrator to discuss the ELI data.	◯	◯	◯	◯
c. I met with other teachers to discuss the ELI data.	◯	◯	◯	◯
d. I met with a mentor teacher to discuss the ELI data.	◯	◯	◯	◯
e. I conducted classroom learning activities informed by ELI data.	◯	◯	◯	◯
f. I used ELI reports to plan whole-group instructional activities.	◯	◯	◯	◯
g. I used ELI reports to plan small group instructional activities.	◯	◯	◯	◯
h. I used ELI reports to plan one-on-one instructional activities.	◯	◯	◯	◯
i. I used ELI reports in combination with other formative assessments.	◯	◯	◯	◯

6. In addition to the ELI, what other student assessments do you use to inform your instruction? Select all that apply.

- Acadience
- aimswebPlus
- FAST
- iStation
- MAP Growth
- mCLASS: DIBELS 8th Edition
- STAR Early Learning
- Other (please specify):_____

7. Did the ELI replace another assessment that you were using? If yes, please provide the name of the assessment that was replaced by the ELI.

- No
- Yes (please specify):_____

8. Did you use data from the ELI report to inform student report cards?

- No
- Yes

9. Please select your level of agreement for each of the following statements. Select one answer for each row.

	Completely disagree	Disagree	Neither agree nor disagree	Agree	Completely agree
a. The ELI seems implementable.	◯	◯	◯	◯	◯
b. The ELI seems possible.	◯	◯	◯	◯	◯
c. The ELI seems doable.	◯	◯	◯	◯	◯
d. The ELI seems easy to use	◯	◯	◯	◯	◯

10. Please indicate the amount of time it took you to complete the following ELI tasks, summing time across <u>all students combined</u>:

	Hours	Minutes
a. Assign ELI ratings.		
b. Enter ELI assessment data into the ELI data dashboard.		
c. Generate ELI reports.		
d. Review reports and data to inform your instruction.		

11. Please describe how your planning and/or instruction changed after attending the ELI training. Describe how this change affected your students. Be specific.

_____

_____

_____

_____

12. Were the following factors useful or not useful when implementing the ELI? Select one answer for each row.

	Not Useful	Useful	Not Applicable
a. The in-person ELI training	◯	◯	◯
b. The ELI webpage	◯	◯	◯
c. The ELI training resources	◯	◯	◯
d. The format of the ELI	◯	◯	◯
e. The ELI data dashboard	◯	◯	◯
f. A school administrator	◯	◯	◯
g. Other assessments I use in kindergarten	◯	◯	◯
h. The ELI reports	◯	◯	◯
i. My professional learning community (PLC)	◯	◯	◯
j. My mentor/instructional coach	◯	◯	◯
k. Planning time during the school day	◯	◯	◯

13. What recommendations do you have for other teachers regarding what is important for *effectively administering* the ELI?

_____

_____

_____

14. What recommendations do you have for other teachers regarding what is important for *effectively using data* from the ELI?

_____

_____

_____

Appendix D: Early Learning Inventory (ELI) Instructions for Coders

Instructions for Coders

1. **Coding conventions**
    a. Include enough context when copying a quote into a node. Copy enough of a quote into a node that will be easily understood by another person reading the quote by itself. If needed, you can include the question asked of the participant.
    b. Look for statements to fit into an existing node, before putting it in an "other" category. A statement may actually belong under a different node in the protocol then where it was asked about.

Early Learning Inventory (ELI) Focus Group Coding Protocol

1. **Using the ELI in Your Classroom**

    a. How were you able to gather the evidence to make ratings?
    b. When did you gather evidence to make ratings?
        i. During planning time/periods
    c. What aspects of the ELI you found ***easy*** to use in your classroom?
        i. Format or navigation through the ELI
        ii. Gathering evidence
        iii. Data presentation
        iv. Other – only use if no other nodes fit
    d. What aspects of the ELI you found ***challenging*** to use in your classroom?
        i. Time
        ii. Amount of information/overwhelming
        iii. Other assessments to complete
        iv. Other challenges– only use if no other nodes fit
    e. Purposes for using the ELI data
    f. ELI data dashboard
        i. Likes
            1. Ease of use
            2. Format choice (of data display)
        ii. Dislikes
    g. ELI reports
        i. Likes
        ii. Dislikes

2. Other Assessments in Your Classroom

a. Other student assessments used while also using the ELI
    i. Acadience
    ii. aimswebPlus
    iii. FAST
    iv. iStation
    v. Literacy First
    vi. MAP
    vii. My Journey
    viii. mCLASS: DIBELS 8th Edition
    ix. STAR Early Learning
    x. Other assessment – only use if no other nodes fit

b. ELI duplication of other assessments/instructional practices
    i. Other assessments align with ELI

3. Local Supports for Data Use

a. Professional Development (PD)
    i. None available
b. Staff in your school or district who can answer your questions about using data.
    i. No one available
    ii. Other teachers
    iii. Interventionists
    iv. Library staff
    v. IT staff
c. Other supports from your school or district that help use the ELI in your classroom more effectively?
    i. More time

4. The ELI Training

d. Activities or resources from the ELI training that were particularly useful.
e. Activities or resources from the training that ended up not being very useful.
f. Any additional training activities or resources that would have been helpful.

5. Recommendations

a. What do you think would make the ELI easier to use?
b. Recommendations for improving the ELI measure.
c. Recommendations for improving the ELI data dashboard

Research Question/Item Crosswalk

Research Questions	Nodes
Research question 1: To what extent do the ELI training and resources prepare teachers to use the ELI?	4
Research question 2: How do teachers report administering the ELI in their classroom and do they report using the ELI data to inform instruction?	1
Research question 3: What are teachers' perceptions about the feasibility and value of using the ELI in their classroom?	1
Research question 4: What are the key facilitators and challenges for teachers implementing the ELI?	1, 2, 3
Research question 5: What could be improved about the ELI training, measure, and technology platform?	1, 4, 5