ESTART VALIDATION STUDY

June 2008

Final Report: Executive Summary

Prepared By:

Mary Maguire Klute, Ph.D.

Independent Consultant

mmklute@gmail.com

Executive Summary

The Early Childhood Standards Assessment of Readiness Tool (ESTART) was developed in 2001 as a part of an effort to create a new set of standards for preschool education in the City and County of Denver. The idea behind the system was to create more uniformity in the meaning of school readiness across early intervention programs, community child care settings, and public preschools, so that public kindergarten classrooms could have similar expectations for the different streams of arriving children. The need for more uniformity was brought to the attention of the city of Denver in the mid 1990's, and in 1998, the early childhood branch of Denver Public Schools (DPS) applied for and received a grant from the U.S. Department of Education's "Goals 2000" program to develop preschool education standards. Although the preschool standards and ESTART were created in Denver, they were based on developmental theory and national Head Start guidelines, and were not intended to be specific to the local community.

The ESTART is intended to inform classroom teachers' instruction and to be a measure of whether and to what extent individual preschool children are benefiting from new standards for early education. During initial item construction, "cross-walks" were conducted with each of these curricula to ensure category and item comprehensiveness, as well as applicability of the ESTART to a number of different types of early childhood settings and curricula. Like the standards, the ESTART covers four content areas (or "domains"): literacy, mathematics, science and motor.

Currently the ESTART is being used by 15 agencies in 31 sites. These sites have a total of 90 classrooms serving 1100 children. While the ESTART has been in use since its development in 2001, this is the first research study undertaken to examine its psychometric properties.

Goals of the Study

The goal of this study is to provide information about the psychometric properties of the ESTART, including its reliability and validity. Reliability and validity are two related but distinct concepts. Reliability has to do with the repeatability or consistency of a measurement.

A measurement is reliable if, when one repeats the same measurement several times, one gets the same, or a quite similar, result. Validity has to do with the extent to which a measure is actually assessing what it designed to assess (i.e., is the ESTART really assessing school readiness?). If a measure is not reliable, it cannot be valid. That is, reliability is considered the upperbound for validity. In practice, the assessment of reliability and validity is not simple. Instead, researchers approach the question in a variety of ways, by asking more specific questions about the psychometric properties of the measure. The challenge is to then try to assemble the answers to these specific questions to draw conclusions about the measure as a whole. The specific questions addressed by the study and the associated psychometric topic are:

1. Are the items included in the ESTART the right items? (Content Validity)
2. How do the items fit together? (Structure)
3. Do items function differently for different subgroups of children? (Bias)

4. Does the ESTART discriminate between different subgroups of children in expectable ways? (Discriminant Validity)
5. Do teachers complete the ESTART similarly after a short period of time elapses? (Test-Retest Reliability)
6. Do two teachers in the same classroom complete the ESTART similarly? (Inter-Rater Reliability)
7. Is the ESTART associated with other established measures of literacy, mathematics, science and motor development? (Concurrent Validity)

This report provides information about each question in turn, including the methods used to address the question (i.e., data collected, analysis strategies, etc.) and the results for each question. This will provide information about the ESTART's strengths and identify areas for improvement, but will not provide a yes or no answer to whether the measure is valid and reliable. In the concluding section of this report, the answers to each specific question are considered together and recommendations for future directions with the ESTART are described.

Key Findings

Overall, results of the study were rather positive, but suggested that revision of the ESTART in a number of areas, particularly the literacy domain is warranted:

Are the items included in the ESTART the right items?

Eleven reviewers provided feedback on the items included in the ESTART. Their reviews suggested that the items included in the ESTART are the right items for the most part. This was especially the case for the math, science and motor domains and less true for the literacy domain. For all domains, the reviewers identified areas for improvement and revision.

How do the ESTART items fit together?

For the mathematics, science, and motor domains, the items generally fit together in the ways suggested by their arrangement on the ESTART score sheet. The literacy domain needed substantial rearrangement. An alternative structure for the literacy domain, which included five subscales, did fit the data. However, these five subscales did not work together as indicators of an overall literacy scale. It is important to note that the analyses of the structure of the ESTART items can only provide ideas about ways to rearrange the items that are already included in the ESTART. These analyses cannot provide any information about what items might be missing. The content validity reviewers' comments can provide insight into this area. It is interesting to note that literacy was the area with the greatest problems with its structure. This fit with the reviewers comments, which were more critical for the literacy domain than for the other three domains. Following some of the advice of the reviewers and adding additional items will likely have implications for the overall structure of the ESTART.

Do items function differently for different subgroups of children?

Analyses were conducted to detect potential bias in ESTART items. Items are biased when children *of the same ability level* score differently on the item depending on what subgroup they belong to (when teachers tend rate children of a particular subgroup lower or higher on an item than children *with the same ability* from another subgroup). A conservative approach was used to identify items that may be biased. As a result, one should not conclude that these items *are* indeed biased. Rather, the identified items should be reconsidered and studied further

The results of these analyses indicated that some items were potentially biased against certain subgroups. Half of the potentially biased items were from the mathematics domain. One might be tempted to simply drop items that are potentially biased from the ESTART. However, if one adopted this approach, the geometry subscale of the mathematics domain would be eliminated entirely.

Does the ESTART discriminate between different groups of children in expectable ways?

We hypothesized three types of group differences. First, we expected that children who spoke primarily English would score higher on the literacy domain of the ESTART than children who spoke primarily Spanish. This hypothesis was generally not supported by the data. This could be due to the fact that the items on the ESTART have more to do with the way a child *uses* language than the actual language itself. Teachers are instructed to think about a child's use of English when completing the ESTART. It may be the case that teachers are not following this instruction when making their ratings. If teachers are thinking about a child's use of *any* language when completing the ESTART, one would expect the pattern of results seen here (i.e., not many differences between the language groups). While taking this approach may give children "credit" for all of their language abilities, this approach would be, by necessity, applied inconsistently because there are classrooms where primarily Spanish-speaking children are taught by teachers who do not speak Spanish. It would be unlikely that monolingual English-speaking teachers would be able to evaluate a child's skill in *any* language as reliably as bilingual teachers would be able to. The situation is further complicated when children speak languages other than English and Spanish. As expected, the mathematics, science and motor domains of the ESTART were not consistently associated with child language.

Our second hypothesis was that scores on the ESTART would increase with child age. This hypothesis was strongly supported by the data in all three rounds and for all four domains of the ESTART. As children age, their ESTART scores increase.

Our third hypothesis was that children with IEPs would score lower on the ESTART than children without IEPs. However, we expected this to correspond with the reason for the IEP (e.g., we expected children with language delays to score lower on the literacy domain of the ESTART, but not necessarily the motor domain). The archival datasets only contained information on whether a child had an IEP, not the reason for it. As a result, with the data available, we were not able to fully test this hypothesis. For the literacy domain, there were fairly consistent associations between literacy scores on the ESTART and IEP status, particularly for phonological sensitivity and book use. For the remaining literacy subscales, as well as

the other domains of the ESTART (mathematics, science and motor), more significant differences were observed in the second year of data collection than the first. Part of this pattern of effects may be due to the difference in the populations for the two years. In 2006-07, the proportion of children with IEPs was much lower than it was in 2005-06. Without information about the reason for an IEP, it is impossible to know if the sample of children in these two years also differed in the reason for their IEP. It is also possible that there was an inconsistency in the way that programs reported this information across the two years (e.g., perhaps including suspected special needs in the first year, but only confirmed IEPs in the second), which may have affected the pattern of results.

In sum, expected differences between groups were observed for age and somewhat for IEP. However, the expected differences by primary language were not observed.

Do teachers complete the ESTART similarly after a short amount of time elapses?

In general, the answer to this question was yes. With a few exceptions, the items and scales of all four domains of the ESTART had good test-retest reliability.

Do two teachers in the same classroom complete the ESTART about the same child similarly?

There was evidence of good interrater reliability for the literacy, mathematics and science domains. However, there were more problems with interrater reliability for the motor domain. Typically, the teachers who participated in the interrater reliability study as second raters were less experienced with the ESTART. The difference in the pattern of results for interrater and test-retest ICCs for motor suggest that perhaps teachers need more support, training, and/or experience to complete this domain reliably.

Is the ESTART associated with other established measures of literacy, mathematics, science and motor development?

The answer to this question was yes, but there was a great deal of association across domains. There was some evidence that some of the cross-domain correlations that were observed may be due to actual clustering of skills in the four domains and not from a flaw in the ESTART. It may be that the ESTART does not do a good job of discriminating between the domains because it is difficult for any measure to do this for this age group, particularly for the three academic domains (literacy, mathematics and science). In response to this dilemma, an approach based on the idea that the ESTART is a measure of overall school readiness was considered. When this approach was adopted, the association between the ESTART scores and the standardized tests were larger and approached the magnitude of correlations desired for establishing criterion validity.

Recommendations

The results of the study definitely indicate that the ESTART has promise. Several directions for further development of the ESTART are suggested by the results:

- *Carefully consider the ideas for revision suggested by the content validity reviewers.* The content validity portion of the study was the only part of the study that provided information about items to

add to the ESTART or specific ideas about how to make adjustments to individual items and their associated rubrics. The reviewers provided thoughtful and thorough feedback. In particular, the comments from the reviewers will make an excellent starting point for revising sections of the ESTART identified for revision by the other parts of the study.

- *Substantially revise the literacy domain of the ESTART before proceeding with any further large-scale study of the ESTART.* It may be particularly useful to pilot test the new literacy domain with a subset of teachers, conducting analyses similar to those conducted in this study after each round of revisions to ensure that the newer version of the literacy domain has improved psychometric properties. Using this approach, one should expect to go through multiple revisions before arriving at the final revised version of the domain.
- *After making revisions to all sections, conduct analyses to test for differential item function.* No data collection aside from teachers completing the ESTART as they normally do is required to accomplish this. The only additional expense would be for time to analyze the data. It is important to assess whether the revisions has the effect of eliminating the differential item function. If not, further revisions will be warranted.
- *Make the assessment of reliability a regular part of the ESTART administration during the revision phase.* If the revisions suggested elsewhere do not alleviate the problems with the reliability of the motor domain, consider making modifications to the training in this area. The motor domain may be an area that is particularly ripe for the use of video clips during training.
- *Delve deeper into the issue of what language teachers should be and are considering when completing the literacy domain of the ESTART.* How to best assess the skills of children who are learning two languages is a thorny issue. It is further exacerbated by the fact that not all developing bilingual children are in classrooms with teachers who speak all of the languages they speak (e.g., many Spanish-speaking children are in classrooms with teachers who do not speak Spanish). Further, the ESTART was developed with English language development in mind and it may not be applicable to development of other languages. At the very least, those who use the ESTART should review the literature on best practices in this area and develop clearer guidelines about how to use the literacy domain of the ESTART with children who are learning multiple languages.
- *Improve demographic data collection to include reasons for IEPs.* Many of the analyses conducted in this study could be repeated annually to monitor the ESTARTs performance during its revision. By improving the demographic data collection to include reasons for IEPs, one would be better positioned to do a better job of examining the ESTARTs discriminant validity.

In sum, the results of this study suggest that with a relatively short revision phase, the ESTART could be much improved. When deciding what revisions to undertake and the time frame for those revisions, one should keep in mind the purposes for which ESTART data will be. Naturally, data that will be used for more high-stakes purposes (e.g., accountability, funding decisions, etc.) need to be gathered with tools that are held to a higher standard that data that are used for more low-stakes purposes (e.g., gathering data continually over time to inform instruction in the classroom). It might be useful to compare the results of this study to the psychometric properties of other tools that are widely accepted to gather data for the same purposes as the ESTART data will be used.