DORA Phonemic Awareness: Validity, Reliability, and Correlation Studies

Sub-test Specifications:

DORA PA assesses students across nine skill areas of phonemic awareness, including: 1) addition, 2) deletion, 3) substitution, 4) identification, 5) categorization, 6) blending, 7) segmenting, 8) isolation, and 9) rhyming. The test is comprised of 50 items, presented either orally or with pictures only, requiring no reading skills. Results are reported in three general categories: Early Phonemic Awareness, which includes word number, rhyming, and identity skills (13 items total); Phoneme Segmentation (15 items total); and Phoneme Manipulation Auditory Analysis Tasks (MAAT), which include blending, addition, deletion, and substitution (22 items total).

DORA PA Reliability:

| DORA PA Overall Person Reliability: .85 |
| DORA PA Cronbach Alpha: .87 |

Introduction

DORA PA is built on an adaptive logic platform, limiting the number of items seen by each student by both grade level skin, which controls the specific items administered according to students’ grade level, and by early termination. Three grade level skins are in place (skin A for students in grades K-2, skin B for students in grades 3-5, or skin C for students in grades 6+). The following table indicates the items administered to each skin:

Table 1

<table>
<thead>
<tr>
<th>DORA PA Question ranges per skin</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Skin</strong></td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>

In addition to the pre-determined questions administered based on the skin appropriate to the grade level of the student, the Let’s Go Learn assessment platform adjusts to the responses of the student, terminating any given cluster of items early when a student has reached non-mastery. There are eight clusters of items. With both the initial skin selection by grade level and the adaptive logic of the assessment, no single student will ever be administered all 50 test items.

In order to evaluate the reliability of DORA PA, the Rasch model was used to evaluate overall model fit for the full assessment, the grade level skins, and the fit of each of the
three reporting groups, as well as to identify any misfitting test items. A sample of 766 students in grades K-12 taken from schools nationwide (23 school districts from Arizona, California, Colorado, Hawaii, Kansas, Missouri, New York, North Carolina, Ohio, and Texas, as well as numerous homeschooling parents) was used, with 499 students in grades K-2 (skin A), 179 students in grades 3-5 (skin B), and 88 students in grades six and above.

**Analysis**

Overall, the assessment appears to be uni-dimensional in nature, with a strong level of person reliability of .85 (as well as a Cronbach alpha of .87). The real person separation level is 2.35, with a model separation level of 2.41. Further, there is virtually no discrepancy between real RMSE (.44) and model RMSE (.43), indicating very few misfit problems.

Each of the grade level skins, when functioning independently, also demonstrates a high level of reliability and a strong fit to the Rasch model. When the items for skin A are examined as an independent model, there are no misfitting items. The overall model fit is still strong (real person reliability = .82, Cronbach alpha = .85, real separation = 2.14).

When the items for skin B are examined as an independent model, the reliability drops slightly (.82), as does the separation (2.18), but still remains overall quite strong, as does the Cronbach alpha (.90).

Finally, when the items for skin C are examined independently, the reliability again drops very slightly (from .85 to .83), as does the separation (from 2.35 to 2.19), still indicating a strong fit. The Cronbach alpha level increases to .9.

Further, the three reported subskills were examined independently as well. When Early Phonemic Awareness was assessed independently, it had a lower level of reliability (.59) and a low level of separation (1.21). There were, however, no misfitting items, and items in the sequence appear to be ordered by difficulty.

In the Segmentation section of the assessment, the person reliability was .60 with a separation of 1.21 but a Cronbach alpha of .85, indicating internal consistency within this sub-section. There are no misfitting items within this section. Items also appear to be sequenced from easiest to hardest within the section.

Finally, in the MAAT section of the assessment, the person reliability is .66, with a separation of 1.41 and a Cronbach alpha of .8.

**Conclusions**

In summary, the DORA PA assessment demonstrates a high level of reliability in the overall assessment model, as well as for all three grade level skins of the assessment. Reliability levels for individual subskills, however, are not as high. For diagnostic purposes and in creating a profile of student strengths and weaknesses, the reliability
levels demonstrated by the individual subskills are still high enough to indicate a confident level of test reliability. It should further be considered, however, that these subskills are not individual testing sections but only reporting sections.

There are a number of explanations as to why the reliability of the subskill sets decreased when they were examined separately. First of all, it is not unexpected to see reliability levels decrease as the number of items assessed decreases. For the skins, we were evaluating between 34 and 45 items, but when the subskills are examined individually, there may be as few as 13 items in the set. When the subskill sets are small, there is less variability in the student responses. As the SD levels rise, the reliability levels fall. Further, there is the issue of missing data, which will affect the reliability levels. Since no one student ever takes all test items, there is always missing data. In the reliability of the overall assessment, the missing data is less of a factor than in the reliability of the individual subskills, due again to the smaller number of items being measured. The missing data there serves to restrict variability even further, in turn increasing the standard error and increasing the mean square error as well.

Test-Retest Study (September 2009):

Let’s Go Learn conducted a test-retest analysis of kindergarten and first grade students \( (n = 98) \) at an elementary school in rural Colorado in September 2009. Grade level deltas were calculated by first taking the absolute value of the change (or delta) between each test and retest sub-test score. Next the mean was calculated from these deltas. All scores are in grade levels, so a grade level delta of .5 means one half of a grade level. SD is standard deviation for mean delta in grade levels.

Overall Phonemic Awareness: Grade Level Delta = .1212, SD = .0952
Early Phonemic Awareness Sub-test: Grade Level Delta = .1645, SD = .1376
Phonemic Segmentation Sub-test: Grade Level Delta = .2500, SD = .2146
MAAT Sub-test: Grade Level Delta = .1756, SD = .1528

Overall grade level deltas are low, indicating a high level of test-retest validity. Further, these tests indicate the appropriateness of DORA PA for use as a pre/post assessment measure.

Correlation Studies: CTOPP and Yopp-Singer

Let’s Go Learn conducted a correlation analysis of kindergarten and first grade students tested within a one-week window using both DORA PA and two nationally recognized assessments of phonemic awareness, the Yopp-Singer Test of Phonemic Segmentation and the Comprehensive Test of Phonological Processing (CTOPP), Phonological Awareness Quotient.

\( DORA PA \) and CTOPP Correlation: .625 \( (n = 52, p < .001) \)
Both of these tests indicate a statistically significant positive correlation with a strong effect size (Cohen suggests using $r$ as effect size with a standard of $r > .5$ indicating strong effect size).

While these correlation scores are lower than might have been expected, it is important to note that there is a significant level of variability introduced into assessing students, particularly at this grade level (Roskos, 2004). Also important to note is the fact that the validity of formative assessments, such as the teacher-administered CTOPP and Yopp-Singer, must take into consideration factors that are not usually considered in test validity, such as personal relevance, articulation of task, and feedback provided by the testing administrator (Johnson & Costello, 2005). Finally, the variability introduced by a teacher-mediated assessment, including non-standardized administration (Wilde et al., 2003), will also contribute to a slightly lower level of correlation. Taking into account all of the additional complexities introduced by assessing much younger children, these levels of correlation do indeed indicate a strong relationship between the testing of DORA PA and other commercially available and validated methods of assessing phonemic awareness.
References


