JSLHR

Research Note

Applying Item Response Theory to the Development of a Screening Adaptation of the Goldman-Fristoe Test of Articulation–Second Edition

Tim Brackenbury,[a] Michael J. Zickar,[a] Benjamin Munson,[b] and Holly L. Storkel[c]

Purpose: Item response theory (IRT) is a psychometric approach to measurement that uses latent trait abilities (e.g., speech sound production skills) to model performance on individual items that vary by difficulty and discrimination. An IRT analysis was applied to preschoolers' productions of the words on the Goldman-Fristoe Test of Articulation–Second Edition (GFTA-2) to identify candidates for a screening measure of speech sound production skills.
Method: The phoneme accuracies from 154 preschoolers, with speech skills on the GFTA-2 ranging from the 1st to above the 90th percentile, were analyzed with a 2-parameter logistic model.
Results: A total of 108 of the 232 phonemes from stimuli in the sounds-in-words subtest fit the IRT model. These phonemes, and subgroups of the most difficult of these phonemes, correlated significantly with the children's overall percentile scores on the GFTA-2. Regression equations calculated for the 5 and 10 most difficult phonemes predicted overall percentile score at levels commensurate with other screening measures.
Conclusions: These results suggest that speech production accuracy can be screened effectively with a small number of sounds. They motivate further research toward the development of a screening measure of children's speech sound production skills whose stimuli consist of a limited number of difficult phonemes.

Screening measures of children's speech sound production skills typically follow the protocol of more comprehensive articulation and phonological tests: Children are asked to name pictures of familiar items that elicit the range of consonant phonemes that are typically acquired during the preschool and early elementary years (e.g., Fluharty, 2001). Pass and fail criteria are based on phonetic transcriptions of children's productions. Screening tasks typically weigh each phoneme or word equally, regardless of the ages at which they are typically developed or their impacts on intelligibility. It is unclear, however, if such a broad-based method is the most effective or efficient way to meet the purposes of screening—namely, identifying the need for further evaluation or referral to another professional (American Speech-Language-Hearing Association, n.d.). It may be that a shorter word list, focused on a subset of phonemes in specific word positions, would be better for distinguishing children with and without potential speech sound disorders (SSD). One challenge to developing an effective screening instrument for SSD is determining which phonemes in words best discriminate children with difficulty from those with typical development. The present study explored this by applying an item response theory (IRT) analysis to 154 children's productions of the sounds-in-words subtest of one commonly used standardized test of children's speech production, the Goldman-Fristoe Test of Articulation–Second Edition (GFTA-2; Goldman & Fristoe, 2000).

IRT is a collection of statistical models that estimate the probability of a person answering an item correctly on the basis of an estimate of the person's underlying latent trait, as well as item parameters that relate to features such as discrimination, difficulty, and guessing. By choosing a particular IRT model, it is possible to better understand how items function, to develop tailored assessments, and to use a wide variety of psychometric tools (see de Ayala, 2009, for basic information on IRT). Although IRT has been primarily used in educational assessment and psychological research, it has been part of research studies in communication sciences and disorders since the 1980s

[a]Bowling Green State University, OH
[b]University of Minnesota, Minneapolis
[c]University of Kansas, Lawrence
Correspondence to Tim Brackenbury: tbracke@bgsu.edu

Editor: Julie Liss
Associate Editor: Tanya Eadie
Received October 7, 2016
Revision received March 13, 2017
Accepted April 11, 2017
https://doi.org/10.1044/2017_JSLHR-L-16-0392

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

(Baylor et al., 2011), with a notable increase in its application over the past 15 years. Recent IRT applications have addressed different aspects of assessment including, but not limited to, examinations of performance differences across populations (Baylor et al., 2013; Baylor et al., 2014; Hula, Doyle, McNeil, & Mikolic, 2006; Justice, Bowles, & Skibbe, 2006) or multiple forms of the same test (Hoffman, Templin, & Rice, 2012); the precision, weighting, or validity of items within a test (Baylor, Yorkson, Bamer, Britton, & Amtmann, 2010; Chenault, Berger, Kremer, & Anteunis, 2013; Edmonds & Donovan, 2012; Fergadiotis, Kellough, & Hula, 2015); and the development of a computerized adaptive version of an existing test (Hula, Kellough, & Fergadiotis, 2015). In addition, and of particular relevance to the present investigation, IRT has successfully assisted the development of screening protocols on the basis of existing tests, banks of test items, and previously collected research data. This work has addressed a wide range of communicative skills, including expressive language skills of Spanish–speaking preschoolers (Guiberson & Rodriguez, 2014); hearing aid acceptance, functionality, and use in adults (Chenault, Anteunis, Kremer, & Berger, 2015; Demorest, Wark, & Erdman, 2011; Mokkink, Knol, van Nispen, & Kramer, 2010); participation across communication contexts by adults with a variety of disorders (Baylor et al., 2013); word naming in adults with aphasia (del Toro et al., 2011); and vocabulary development in young children (Makransky, Dale, Havmose, & Bleses, 2016).

The current study focused on childhood SSD. Children with SSD present with poor speech intelligibility as a result of motoric, linguistic, cognitive, sensory, or unspecified issues. Estimates of their prevalence among preschool- and elementary-aged children range from 2% to 25% of the general population (Law, Boyle, Harris, Harkness, & Nye, 2000). Clinicians and researchers identify children with SSD through a combination of standardized tests, spontaneous speech samples, and measures to rule out other causes, such as an oral mechanism examination to rule out structural anomalies. To date, there have been no published studies examining the use of IRT to develop an assessment tool for children with SSD. The present study addresses this need through the following research questions. The first focuses on the phonemes identified with the IRT model. The second and third explore the utility of those phonemes, and subsets of the phonemes with the greatest difficulty scores, to serve as a screening measure of children's speech sound production skills.

1. Which phonemes within the stimuli of the sounds-in-words subtest of the GFTA-2 will fit within an IRT model?
2. How well do children's performance on the phonemes in the IRT model, and subsets of those phonemes, correlate with their percentile score performance on the GFTA-2?
3. How strongly can children's percentile score performance on the GFTA-2 and identification as having or not having SSD be predicted from their performance on the phonemes from the IRT model and subsets of those phonemes?

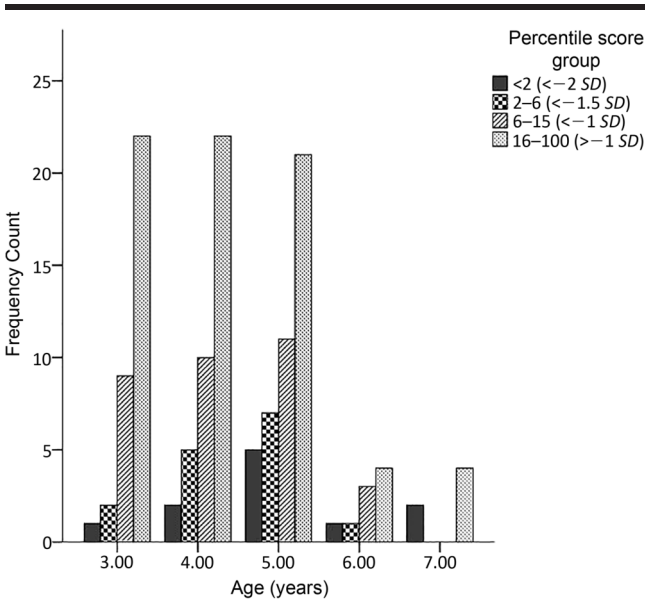
Method

The participants were 154 monolingual boys and girls between 3 and 7 years of age, with and without SSD. This age group was selected because this is an age at which SSD is most likely to be diagnosed, and hence which is subject to a high number of speech and language screening assessments. The participants' data were collected as part of multiple previous research studies conducted by the third and fourth authors (e.g., Munson, Baylis, Krause, & Yim, 2010; Munson & Krause, 2017; Storkel & Hoover, 2010; Storkel, Maekawa, & Hoover, 2010). All usable data were included; no potential participants were specifically included or excluded in order to best fit the IRT model. International Phonetic Alphabet transcriptions were available for each child's productions of the 53 words on the sounds-in-words subtest of the GFTA-2. Age and percentile scores, however, were only available for 133 of the participants. These children had a mean age of 57.2 months (4 years, 9 months, ± 12.73 months) and included 34 three-year-olds, 39 four-year-olds, 44 five-year-olds, 9 six-year-olds, and 6 seven-year-olds (one child's age was not identified).

The GFTA-2 (Goldman & Fristoe, 2000) was chosen because it is among the most widely used standardized tests of children's speech production conducted and used in the 15 years prior to this study. Its norming sample includes children with and without SSD between the ages of 2 years, 0 months and 21 years, 11 months. Standard scores on the GFTA-2 are based on children's performances on the sounds-in-words subtest, in which their productions of target phonemes within a picture-naming task are scored as correct or incorrect. The GFTA-2 percentile score performances of the 133 participants with complete data sets ranged from 1 to 98, with a mean of 32.92 (± 29.55). The children with SSD included articulatory and phonological issues of unknown origin, not secondary to other sensory or cognitive issues or diagnoses such as childhood apraxia of speech. As shown in Figure 1, the percentile score performances of these children, at each age, reflected the GFTA-2's distribution in which progressively fewer children scored at lower ends of the percentile score range. The GFTA-2 was administered and scored using its standard method, in which children are prompted to name pictures. Children who do not name pictures spontaneously are given a series of progressively greater support until they produce the target word. Responses are phonetically transcribed.

Each of the 154 participants' attempts at the 232 individual phonemes included in the GFTA-2's sounds-in-words subtest was treated as a separate item and scored dichotomously as correct or incorrect. Because phonemes were nested within individual words (e.g., the /s/ in *house* was discrete from the /s/ in *stars*), consonant clusters were categorized by their constituent phonemes (e.g., *stars* included separate entries for /s/, /t/, /ɑ/, /r/, and /z/). Items

Figure 1. Histogram depicting Goldman-Fristoe Test of Articulation–Second Edition performance by age and standard deviation score.



that were answered incorrectly by fewer than five participants were eliminated because they did not have enough variance for analysis. IRTPro 2.0 was used to fit the two-parameter logistic (2PL) model to all 232 items of the test (Paek & Han, 2012). IRTPro is a software package that estimates IRT parameters using a variety of possible IRT models. Although there are many models that could have been chosen for these data, the 2PL was selected because it allows items to vary on both difficulty and discrimination, two features found to be important in modeling items. In addition, the 2PL was a reasonable choice given the relatively small sample size. The 2PL model is represented by the following formula:

$$P(u = 1|\theta) = \frac{1}{1 + e^{(-a(\theta - b))}}$$

where a refers to the item *discrimination* (i.e., the strength of the relation of that item to the underlying trait), b refers to the item *difficulty,* and θ refers to the latent trait being measured by the trait. The 2PL formula uses the item parameters (a and b) in conjunction with the person parameter (θ) to predict the probability of answering an item u correctly. The parameters, both item and person, are estimated via IRTPro using maximum likelihood estimation. An iterative process was carried out to estimate item parameters, eliminating items that did not fit the 2PL through using the χ^2 goodness-of-fit statistics estimated by IRTPro. After eliminating poor fit items, the analysis was re-run, continuing to throw out items until the model fit acceptably well (i.e., there were no items that had significant misfit as judged by IRTPro's χ^2 statistics).

Once the IRT analysis was complete, additional statistical analyses were conducted to identify (a) the degree to which the model accounted for the children's overall performances on the GFTA-2, and (b) the predictive accuracy of a subset of phonemes from the model to discriminate children with and without potential SSD.

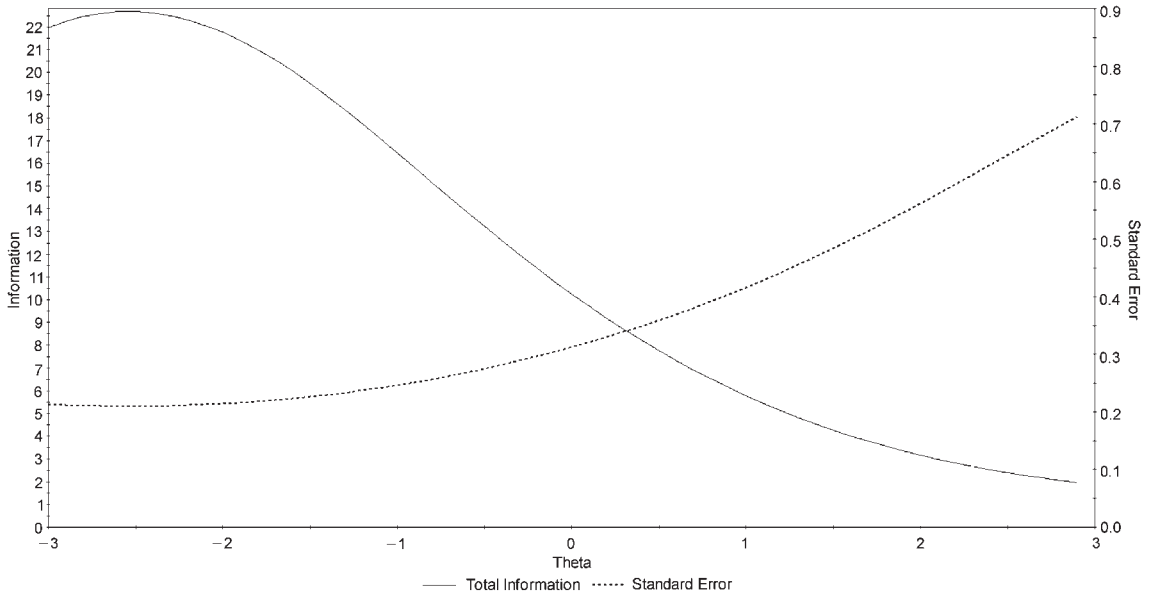
Results

To test whether the data satisfied the requirement of sufficient unidimensionality, a factor analysis of tetrachoric interitem correlations was conducted (necessary because the data were dichotomous) and found that the first factor accounted for 31.5% of the variance in the scale. This satisfied the requirement that Reckase (1979) identified that the first factor in an exploratory factor analysis needed to account for at least 25% of the variance to satisfy the unidimensionality assumption of IRT.

The final IRT model consisted of 108 phonemes, which are listed in the Appendix in order from the highest to lowest difficulty score. They included all of the American English consonants, except for /h/ and /ʒ/, and the vowels /i, ɪ, ɛ, æ, ə, ʌ, ɚ, aɪ, aʊ/. The consonants, as a group, occurred in initial and final syllable positions, and as singletons and within clusters. The phonemes in the 2PL model were from 49 of the 53 words on the GFTA-2 (i.e., all words except for *ball, house, ring,* and *thumb*) and included 47 of the 92 phonemes used to determine percentile scores on the GFTA-2.

The analysis began by exploring which area of the underlying trait, commonly denoted by the Greek letter θ in IRT research, provided the most psychometric information. Information is an IRT-based concept that quantifies the amount of precision provided by the test at varying levels of θ. Traditional measurements of precision, such as standard error of measurement or reliability, assume that the precision is uniform throughout the range of the trait being measured. This assumption is likely untrue for many tests given that some tests are designed to be easy so that at-risk individuals can be identified, whereas other tests are designed to identify top talent. In IRT, the *test information function* allows test users to identify at what range of the trait the test provides precision and at what ranges the test is relatively imprecise. Figure 2 shows the test information function for the 108-item 2PL model. The most information was provided in the negative range of the trait continuum, as demonstrated on the left side of Figure 2 in which the total information values were higher than the standard error, meaning that the test as a whole was able to provide the most precise measurement at the low ability. There was relatively little precision at the high end, as shown on the right side of Figure 2 where the standard error outranked the total information. This suggests that an instrument based on this model would not be able to distinguish well between children with the very best speech sound production skills from those at the upper end of the normal range. The greater precision in the negative range is in line with the goal of using items from the GFTA-2 as a screening test because it

Figure 2. Test information function for the 108-item two-parameter logistic model.



emphasizes differentiating children who are functioning below the normal range from those who are within the normal range. To increase measurement precision in the positive range, it would be necessary to write additional items that were high in difficulty and able to discriminate between average and high ability respondents.

Correlations between the percentile scores on the GFTA-2 from the 133 participants with complete data sets and their summed accuracy scores of (a) the 92 phonemes used to determine the percentile scores, (b) the 108 phonemes in the 2PL model, and (c) various subsets of the phonemes in the model with the greatest difficulty scores are presented in Table 1. Significant correlations with GFTA-2 percentile scores were found for each of the groups assessed, with r^2 values from .16 to .66. Two sets of multiple regressions were run to determine how well combinations of the percentile score phonemes and each group of 2PL phonemes contributed to the children's GFTA-2 percentile scores. In the first set, the percentile phonemes were entered prior to the 2PL phonemes. As shown in Table 1, all but two of the 2PL groups (the 108 and the one phoneme 2PL groups) contributed significantly after the effects of the percentile phonemes were accounted for ($p \leq .05$). In the second set, the 2PL phonemes were entered before the percentile score phonemes. In this set, the percentile phonemes accounted for significant additional variance

Table 1. Multiple stepwise regressions for two sets of predictors, percentile score phonemes, and sets of phonemes from the two-parameter logistic (2PL) model on 133 participants' percentile scores on the Goldman-Fristoe Test of Articulation–Second Edition.

Comparison 2PL phoneme group	Percentile score phonemes entered before 2PL phonemes				2PL phonemes entered before percentile score phonemes			
	Run	*r*	r^2	*p*	Run	*r*	r^2	*p*
108	1	.71	.50	<.01	1	.70	.49	<.01
	2			.75	2	.71	.50	<.01
20 most difficult	1	.71	.50	<.01	1	.76	.57	<.01
	2	.76	.57	<.01	2			.66
15 most difficult	1	.71	.50	<.01	1	.76	.58	<.01
	2	.76	.58	<.01	2			.79
10 most difficult	1	.71	.50	<.01	1	.76	.57	<.01
	2	.76	.58	<.01	2			.22
5 most difficult	1	.71	.50	<.01	1	.73	.53	<.01
	2	.74	.55	<.01	2	.74	.55	<.01
3 most difficult	1	.71	.50	<.01	1	.68	.46	<.01
	2	.73	.53	<.01	2	.73	.53	<.01
2 most difficult	1	.71	.50	<.01	1	.66	.43	<.01
	2	.72	.52	<.01	2	.72	.52	<.01
1 most difficult	1	.71	.50	<.01	1	.42	.18	<.01
	2			.13	2	.71	.51	<.01

after the 1-, 2-, 3-, and 5-phoneme 2PL groups were entered ($p < .01$). However, the contribution from the percentile phoneme group was not significant for the 2PL groups with 10, 15, and 20 phonemes ($p > .22$).

The 3-, 5-, and 10-phoneme 2PL groups (see Table 2) were examined as potential candidates for a screening measure because they were the smallest 2PL groups that accounted for as much variability in GFTA-2 percentile scores as the 92 phonemes used to determine those scores ($r^2 = .53, .62, .67$, and $.57$, respectively). This process began by calculating separate regression equations for each of these groups on the children's GFTA-2 percentile scores. All three equations were significant at the .05 level ($p < .01$). The regression equation for the 3-phoneme group was predicted GFTA-2 percentile score = (/s/ in *stars* × 12.66) + (/r/ in *crying* × 29.76) + (/θ/ in *bath* × 12.54) + 7.94. The 5-phoneme group's regression equation was predicted GFTA-2 percentile score = (/s/ in *stars* × 10.38) + (/r/ in *crying* × 21.53) + (/θ/ in *bath* × 12.61) + (/r/ in *tree* × 5.11) + (/ʃ/ in *fishing* × 13.61) + 1.81. Last, the 10-phoneme group's regression equation was predicted GFTA-2 percentile score = (/s/ in *stars* × 6.42) + (/r/ in *crying* × 6.40) + (/θ/ in *bath* × 6.05) + (/r/ in *tree* × −1.19) + (/ʃ/ in *fishing* × 12.21) + (/r/ in *brush* × 7.35) + (/ð/ in *feather* × 9.15) + (/ŋ/ in *monkey* × 9.14) + (/r/ in *rabbit* × 16.08) + (/v/ in *vacuum* × −2.35) − 0.65.

Each of these equations was then applied to the 133 participants' productions, yielding estimated percentile scores. Their utilities as speech screening measures were evaluated by calculating sensitivity, specificity, and likelihood ratios for cut-off points that best approximated 1 *SD* below the mean. In general, sensitivity and specificity scores ≥ 80%, positive likelihood ratios ≥ 3, and negative likelihood ratios ≤ 0.3 are considered preferable (Dollaghan, 2007). As shown in Table 3, the 3-phoneme regression equation was better at accurately identifying children performing within the average range than those below (sensitivity = 62% and specificity = 76%). The 5- and 10-phoneme regression equations outperformed the 3-phoneme equation, and showed the opposite pattern (with sensitivities at 84% and 88%, and specificities at 74% and 70%, respectively). The likelihood ratio results also favored the 5- and 10-phoneme regression equations. The positive likelihood ratios were similar for all three equations, between 2.56 and 3.26, indicating small-to-moderate probabilities that the children below the cut-off score truly had SSD (Dollaghan, 2007). The negative likelihood ratio of 0.50 for the 3-phoneme equation yielded a mild probability, whereas the 0.21 and 0.16 results for the 5- and 10-phoneme equations, respectively, indicated stronger probabilities that children scoring above the cut off did not have SSD (Dollaghan, 2007). To determine if other phonemes within the 2PL model would be more accurate, successive blocks of the next 5 and 10 most discriminating phonemes across the entire model were run using the same process. The results for all of these calculations were similar to those above, with sensitivity scores consistently 20% or more lower than specificity scores for the same phonemes.

Table 2. Predicted Goldman-Fristoe Test of Articulation–Second Edition regression equations for the 3-, 5-, and 10-phoneme groups.

3-phoneme group		5-phoneme group		10-phoneme group	
/s/ in *stars*	× 12.66	/s/ in *stars*	× 10.38	/s/ in *stars*	× 6.42
/r/ in crying	× 29.76	/r/ in crying	× 21.53	/r/ in crying	× 6.40
/θ/ in bath	× 12.54	/θ/ in bath	× 12.61	/θ/ in bath	× 6.05
+	7.94	/r/ in tree	× 5.11	/r/ in tree	× −1.19
Predicted percentile score		/ʃ/ in fishing	× 13.61	/ʃ/ in fishing	× 12.21
		+	1.81	/r/ in brush	× 7.35
		Predicted percentile score		/ð/ in feather	× 9.15
				/ŋ/ in monkey	× 9.14
				/r/ in rabbit	× 16.08
				/v/ in vacuum	× −2.35
				+	−0.65
				Predicted percentile score	

Discussion

The 2PL model identified 108 phonemes from the stimuli in the sounds-in-words subtest of the GFTA-2 that significantly discriminated performance for preschool- and early elementary-aged children. These included the majority of consonants and vowels in American English, but did not strongly overlap with the phonemes used by the GFTA-2 to determine percentile scores. This is not surprising, as the test was "designed to provide a controlled sample of a child's spontaneous production in words *of the most frequently occurring consonant sounds in Standard American English* [emphasis added]" (Goldman & Fristoe, 2000, p. 7). In other words, the phonemes assessed by the GFTA-2 were chosen to represent the wide range of consonant sounds, not by how well they discriminated performance. In addition, the GFTA-2 scoring system weighs each phoneme equally, despite the variations in ages at which they are typically developed or their impacts on intelligibility. These features are similar to other tests based on classical test theory (e.g., deVellis, 2006).

In contrast, the 2PL phonemes and their regression equations align more closely with IRT (e.g., Embretson & Reise, 2000) because they include only the phonemes with the greatest difficulty scores, and each phoneme is individually weighted on the basis of its impact on the predicted score. The phonemes that occurred within the 10 most difficult items of the 2PL model were /s, r, θ, ʃ, ð, ŋ, v/. All of these except for /ŋ/, depending on the data source, are typically later-developing phonemes in American English (e.g., Smit, Hand, Freilinger, Bernthal, & Bird, 1990). The 10 most difficult words also included the target phonemes in the challenging contexts of consonant clusters, medial positions of multisyllabic words, and word final position. It is likely that these aspects of the target phonemes' difficulty are what contributed to their potential as screening items, and not simply their inclusion within the GFTA-2's stimuli. Further, the full 2PL model's inclusion of both easy and difficult phonemes may explain why it was more

Table 3. Sensitivity, specificity, and likelihood ratio calculations on the basis of three different regression equations developed from the two-parameter logistic model, at a cut off of the 16th percentile on the Goldman-Fristoe Test of Articulation–Second Edition.

Regression equation	Regression cut-off score applied	Sensitivity (%)	Specificity (%)	Positive likelihood ratio	Negative likelihood ratio
3-phoneme	20	62	76	2.56	0.50
5-phoneme	17	84	74	3.26	0.21
10-phoneme	17	88	70	2.98	0.16

precise at discriminating performance at the lower end of the spectrum than the higher end. A measure that consists of only difficult phonemes may be better at discriminating performance across the spectrum. Taken together, these results suggest that future screening measures of children's speech sound production skills, whether they are or are not developed from existing tests, should consider stimuli that include difficult phonemes in challenging contexts.

The predictive abilities of the 3-, 5-, and 10-phoneme groups were in a positive, but not overwhelming, direction. As a group, however, they were within the ranges reported for other assessments of child speech and language disorders. Two systematic reviews of screening measures of preschoolers' speech and language skills (Law et al., 2000; Nelson, Nygern, Walker, & Panoscha, 2006), for example, revealed sensitivity ranges from 17%–100%, and specificity ranges from 14%–100%. It is noted, however, that approximately half of these screening measures identified fell below the suggested 80% lower limits for sensitivity or specificity (Dollaghan, 2007). Of the three groups assessed in this study, the one with 3 phonemes appeared to be the weakest, due to its poorer sensitivity, specificity, and likelihood values. The results for the 5- and 10-phoneme groups were both better and fairly similar to each other. Caution is advised before directly applying the results of this study to clinical or research settings. Because the regression equations were calculated from a subset of children used to develop the 2PL model, for example, it is currently unclear how well these results will generalize to other children. In addition, the concurrent validity of the 5- and 10- phoneme groups with other standardized measures of speech sound production should be evaluated. As a result, direct applications of the phonemes and words within the 2PL model to speech screening are not recommended without additional research.

Although successful, the 2PL model was relatively simple, due to its dichotomous scoring of item responses. With larger data sets, more flexible models could be used to fit these data, including the three-parameter logistic model that allows for guessing, and polytomous IRT models that would allow graded responses to be scored (see Zickar, 2002). The latter might be useful in determining whether a scoring system that addresses the specific type of errors (such as phonological process or distinctive feature differences) would help improve the measurement. In addition, larger sample sizes would allow for us to estimate these more complex models as well as model some of the easy items that few children answered incorrectly. Additional areas for future exploration on this topic include comparing the results of similar IRT analyses on other measures of speech sound production and examining if and how the IRT results may vary across age groups.

Acknowledgments

The data analyzed in this article were collected in studies that were funded by a National Institute of Health Grant R03 DC005702 to Benjamin Munson and Grant R03 DC006545 to Holly L. Storkel.

References

American Speech-Language-Hearing Association (n.d.). *Speech sound disorders: Articulation and phonology.* (Practice Portal). Retrieved from www.asha.org/Practice-Portal/Clinical-Topics/Articulation-and-Phonology

Baylor, C., Hula, W., Donovan, N. J., Doyle, P. J., Kendall, D., & Yorkson, K. (2011). An introduction to item response theory and Rasch models for speech-language pathologists. *American Journal of Speech-Language Pathology, 20,* 243–259.

Baylor, C., McAuliffe, M. J., Hughes, L. E., Yorkston, K., Anderson, T., Kim, J., & Amtmann, D. (2014). A differential item functioning (DIF) analysis of the Communicative Participation Item Bank (CPIB): Comparing individuals with Parkinson's disease from the United States and New Zealand. *Journal of Speech, Language, and Hearing Research, 57,* 90–95.

Baylor, C., Yorkston, K., Bamer, A., Britton, D., & Amtmann, D. (2010). Variables associated with communicative participation in people with multiple sclerosis: A regression analysis. *American Journal of Speech-Language Pathology, 19,* 143–153.

Baylor, C., Yorkston, K., Eadie, T., Kim, J., Chung, H., & Amtmann, D. (2013). The Communicative Participation Item Bank (CPIB): Item bank calibration and development of a disorder-generic short form. *Journal of Speech, Language, and Hearing Research, 56,* 1190–1208.

Chenault, M., Anteunis, L., Kremer, B., & Berger, M. (2015). An investigation of measurement equivalence in hearing response scales: Refinement of a questionnaire for use in hearing screening. *American Journal of Audiology, 24,* 188–203.

Chenault, M., Berger, M., Kremer, B., & Anteunis, L. (2013). Quantification of experienced hearing problems with item response theory. *American Journal of Audiology, 22,* 252–262.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford.

del Toro, C. M., Bislick, L. P., Comer, M., Velozo, C., Romero, S., Gonzalez Rothi, L. J., & Kendall, D. L. (2011). Development of a short form of the Boston Naming Test for individuals with aphasia. *Journal of Speech, Language, and Hearing Research, 54,* 1089–1100.

Demorest, M. E., Wark, D. J., & Erdman, S. A. (2011). Development of the screening test for hearing problems. *American Journal of Audiology, 20,* 100–110.

deVellis, R. F. (2006). Classical test theory. *Medical Care, 44,* S50–S59.

Dollaghan, C. A. (2007). *The handbook for evidence-based practice in communication disorders*. Baltimore, MD: Brookes.

Edmonds, L. A., & Donovan, N. J. (2012). Item-level psychometrics and predictors of performance for Spanish/English bilingual speakers on an object and action naming battery. *Journal of Speech, Language, and Hearing Research, 55,* 359–381.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fergadiotis, G., Kellough, S., & Hula, W. D. (2015). Item response theory modeling of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research, 58,* 865–877.

Fluharty, N. B. (2001). *Fluharty Preschool Speech and Language Screening Test*. Austin, TX: Pro-Ed.

Goldman, R., & Fristoe, M. (2000). *Goldman-Fristoe Test of Articulation–Second Edition*. San Antonio, TX: Pearson.

Guiberson, M., & Rodriguez, B. L. (2014). Rasch analysis of a Spanish language-screening parent survey. *Research in Developmental Disabilities, 35,* 646–656.

Hoffman, L., Templin, J., & Rice, M. L. (2012). Linking outcomes from Peabody Picture Vocabulary Test forms using item response models. *Journal of Speech, Language, and Hearing Research, 55,* 754–763.

Hula, W., Doyle, P. J., McNeil, M. R., & Mikolic, J. M. (2006). Rasch modeling of Revised Token Test performance: Validity and sensitivity to change. *Journal of Speech, Language, and Hearing, Research, 49,* 27–46.

Hula, W., Kellough, S., & Fergadiotis, G. (2015). Development and simulation testing of a computerized adaptive version of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research, 58,* 878–890.

Justice, L. M., Bowles, R. P., & Skibbe, L. E. (2006). Measuring preschool attainment of print-concept knowledge: A study of typical and at-risk 3- to 5-year-old children using item response theory. *Language, Speech, and Hearing Services in Schools, 37,* 224–235.

Law, J., Boyle, J., Harris, F., Harkness, A. & Nye, C. (2000). The feasibility of universal screening for primary speech and language delay: Findings from a systematic review of the literature. *Developmental Medicine and Child Neurology, 42,* 190–200.

Makransky, G., Dale, P. S., Havmose, P., & Bleses, D. (2016). An item response theory–based, computerized adaptive testing version of the MacArthur–Bates Communicative Development Inventory: Words & Sentences (CDI:WS). *Journal of Speech, Language, and Hearing Research, 59,* 281–289.

Mokkink, L. B., Knol, D. L., van Nispen, R. M. A., & Kramer, S. E. (2010). Improving the quality and applicability of the Dutch scales of the Communication Profile for the Hearing Impaired using item response theory. *Journal of Speech, Language, and Hearing Research, 53,* 556–571.

Munson, B., Baylis, A. L., Krause, M. O. & Yim, D.-S. (2010). Representation and access in phonological impairment. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10*. Berlin, Germany: Walter de Gruyter.

Munson, B., & Krause, M. O. P. (2017). Phonological encoding in speech sound disorder: Evidence from a cross-modal priming experiment. *International Journal of Language and Communication Disorders, 52,* 282–300. https://doi.org/10.1111/1460-6984.12271

Nelson, H. D., Nygren, P., Walker, M., & Panoscha, R. (2006). Screening for speech and language delay in preschool children: Systematic evidence review for the US Preventive Services Task Force. *Pediatrics, 117,* e298–e319.

Paek, I., & Han, K. T. (2012). IRTPRO 2.1 for Windows (Item response theory for patient-reported outcomes). *Applied Psychological Measurement, 37,* 242–252.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics, 4,* 207–230.

Smit, A. B., Hand, L., Freilinger, J., Bernthal, J., & Bird, A. (1990). The Iowa Articulation Norms Project and its Nebraska replication. *Journal of Speech and Hearing Disorders, 55,* 779–798.

Storkel, H. L., & Hoover, J. R. (2010). Word learning by children with phonological delays: Differentiating effects of phonotactic probability and neighborhood density. *Journal of Communication Disorders, 43,* 105–119.

Storkel, H. L., Maekawa, J., & Hoover, J. R. (2010). Differentiating the effects of phonotactic probability and neighborhood density on vocabulary comprehension and production: A comparison of preschool children with versus without phonological delays. *Journal of Speech, Language, and Hearing Research, 53,* 933–949.

Zickar, M. J. (2002). Modeling data with polytomous item response theory. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 123–155). San Francisco, CA: Jossey-Bass.

Appendix

The 108 phonemes from the two-parameter logistic (2PL) model, ranked from highest to lowest difficulty score.

Phoneme	Difficulty score	Phoneme	Difficulty score
/s/ in stars*	1.21	/k/ in duck*	–0.12
/r/ in crying*	1.15	/n/ in pencils	–0.13
/θ/ in bath*	1.15	/tʃ/ in watch*	–0.13
/r/ in tree*	1.03	/f/ in feather	–0.15
/ʃ/ in fishing*	0.92	/k/ in vacuum	–0.16
/r/ in brush*	0.61	/l/ in lamp*	–0.17
/ð/ in feather*	0.54	/f/ in knife*	–0.21
/ŋ/ in monkey	0.48	/ʌ/ in balloons	–0.24
/r/ in rabbit*	0.47	/w/ in watches	–0.33
/v/ in vacuum*	0.43	/ə/ in wagon	–0.34
/r/ in green*	0.41	/n/ in orange	–0.36
/r/ in frog	0.4	/ɛ/ in feather	–0.41
/z/ in zipper*	0.4	/ɛ/ in pencils	–0.41
/g/ in wagon*	0.4	/ɪ/ in orange	–0.41
/k/ in crying*	0.37	/t/ in telephone*	–0.41
/ʃ/ in shovel*	0.34	/d/ in drum*	–0.43
/ð/ in this*	0.34	/r/ in chair	–0.44
/θ/ in bathtub*	0.33	/ə/ in pajamas	–0.46
/z/ in glasses	0.32	/n/ in telephone	–0.48
/s/ in swimming*	0.31	/ə/ in pencils	–0.55
/ɪ/ in watches	0.27	/ɪ/ in scissors	–0.59
/dʒ/ in pajamas*	0.26	/m/ in vacuum	–0.6
/f/ in five	0.24	/n/ in window	–0.7
/ɪ/ in finger	0.24	/t/ in bathtub*	–0.76
/l/ in balloons*	0.22	/æ/ in glasses	–0.85
/l/ in flowers*	0.2	/u/ in vacuum	–0.88
first /k/ in quack*	0.19	/d/ in window*	–0.88
/f/ in fishing*	0.19	/aʊ/ in flowers	–0.91
/ʌ/ in pajamas	0.19	/b/ in rabbit*	–1.06
/ə˞/ in girl	0.18	/u/ in spoon	–1.07
/p/ in pajamas	0.17	/ɪ/ fishing	–1.14
/g/ in girl*	0.16	/aɪ/ in slide	–1.16
/ŋ/ in finger*	0.16	/æ/ in bathtub	–1.2
/l/ in yellow	0.16	/ɪ/ in window	–1.22
/ə˞/ in finger	0.14	/æ/ in lamp	–1.26
/l/ in telephone	0.12	/n/ balloons	–1.3
/r/ in orange	0.11	/æ/ in bath	–1.56
/l/ in glasses*	0.11	/æ/ in rabbit	–1.65
/k/ in clown*	0.1	/ʌ/ in bathtub	–1.8
/l/ in shovel	0.1	/ʌ/ in shovel	–1.88
/ə˞/ in scissors	0.08	/i/ in green	–1.98
first /z/ in scissors*	0.08	/n/ in wagon	–2.46
/k/ in cup*	0.06	/b/ in blue*	–3.16
/k/ in car	0.06	/t/ stars*	–3.56
/l/ in plane*	0.06	/n/ in knife*	–3.91
/z/ in pajamas	0.05	/ə/ in banana	–4.5
/r/ in carrot*	0.05		
/ʌ/ in banana	0.04		
/ə˞/ in feather	0.04		
/dʒ/ in orange*	0.03		
/r/ in stars	0.03		
/j/ in vacuum	0.02		
/g/ in finger	0		
/dʒ/ in jumping*	–0.02		
/aɪ/ in crying	–0.04		
/b/ in banana	–0.04		
/v/ in shovel*	–0.05		
/f/ in finger	–0.05		
/j/ in yellow*	–0.06		
/w/ in swimming*	–0.07		
/ŋ/ in fishing	–0.08		
/ŋ/ in jumping	–0.12		

*Phonemes included in both the 2PL model and the percentile scoring for the Goldman-Fristoe Test of Articulation–Second Edition.