Applying Item Response Theory to the Development of a Screening Adaptation of the Goldman-Fristoe Test of Articulation-2

Tim Brackenbury, Michael Zickar — Bowling Green State University; Benjamin Munson; Holly Storkel — KU

Introduction

Screening measures of children's speech sound production skills sample a variety of phonemes in words, regardless of their typical ages of acquisition or impacts on identification. This study examined if Item Response Theory (IRT) could be used to identify phonemes for a screening protocol based on their abilities to discriminate children with and without speech sound disorders. An IRT analysis was applied to 154 children's productions of the sounds-in-words subtest of the Goldman-Fristoe Test of Articulation – 2 (GFTA-2; Goldman & Fristoe, 2000).

Item Response Theory (IRT)

A psychometric approach to measurement that uses latent trait abilities to model performance, based on individual items that vary by difficulty and discrimination.

Applied for a variety of assessment purposes in Communication Sciences and Disorders:

- Comparing performance differences across populations
- Comparing multiple forms of the same test
- Validating individual items within a test
- Creating tests that are adaptive to individuals' responses
- Developing screening protocols
  - Spanish-speaking preschoolers
  - Hearing aid use by adults
  - Word finding in adults with aphasia
  - Vocabulary development in young children

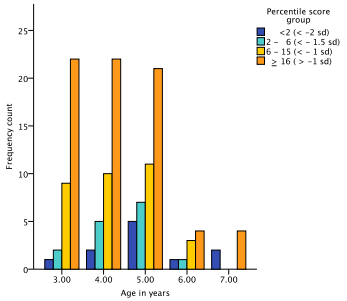
Data and Participants

I. GFTA-2 Transcriptions

- 154 boys and girls between 3 and 7 years of age
- with and without speech sound disorders
  - articulatory and/or phonological issues of unknown origin
- no identified sensory issues, cognitive problems, or other developmental delays
- IPA transcriptions of all 53 words of the sounds-in-words subtest of the GFTA-2, 232 individual phonemes
  - data collected as part of other research studies

II. GFTA-2 Percentile Scores

- from a subset of the 154 children
- 133 boys and girls between 3 and 7 years of age
- with and without speech sound disorders
- IPA transcriptions, age, and percentile score data



Analysis

Each participant's productions of the 232 phonemes on the GFTA-2 were scored dichotomously as correct or incorrect. Phonemes answered incorrectly by fewer than 5 participants were discarded, due to a lack of variance (n = 16).

IRTPro2.0 (Paek & Han, 2012) was used to fit a two-parameter logistic model (2PL): difficulty and discrimination.

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

a = item difficulty; b = item discrimination; θ = latent trait, u = individual item

Iterative process eliminating items that did not fit the 2PL through using the $\chi 2$ goodness-of-fit statistics estimated by IRTPro.

I. Which phonemes from the sounds-in-words subtest of the GFTA-2 fit the IRT model?

The final 2PL model included 106 phonemes:

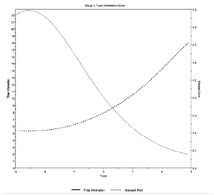
all consonants except for /h, ʒ/ (/ʒ/ is not in the GFTA-2 stimuli)

the vowels /i, ɪ, ɛ, æ, ə, ʌ, ɚ-, aɪ, aʊ/

These came from 49 of the 53 words on the GFTA-2.

They included 46 of the 92 phonemes used to calculate percentile scores on the GFTA-2.

The IRT model provided the most precise information at the lower end of the trait continuum, suggesting it would be better for identifying children with speech sound disorders than "gifted" speakers.



II. How well did the participants' accuracies on the phonemes in the IRT model, and subsets of those phonemes, correlate with their percentile scores on the GFTA-2?

Phoneme Group	r	r^2	p
92 percentile score phonemes	0.71	0.50	< 0.01
108 2PL phonemes	0.70	0.49	< 0.01
20 most difficult 2PL phonemes	0.76	0.57	< 0.01
15 most difficult 2PL phonemes	0.76	0.58	< 0.01
10 most difficult 2PL phonemes	0.76	0.57	< 0.01
5 most difficult 2PL phonemes	0.73	0.53	< 0.01
3 most difficult 2PL phonemes	0.68	0.46	< 0.01
2 most difficult 2PL phonemes	0.66	0.43	< 0.01
1 most difficult 2PL phonemes	0.42	0.18	< 0.01

III. How strongly did the participants' accuracies on subsets of the IRT model predict their speech status?

The 3-, 5-, and 10-phoneme 2PL groups were examined because they accounted for as much variability in GFTA-2 percentile scores as the 92 phonemes used to determine those scores. Results from separate multiple regressions. Each significant ($p < 0.01$).

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

The utilities of these equations as screening measures were evaluated based on the predicted and actual performances of 133 participants.

Measure	3-Phoneme Group	5-Phoneme Group	10-Phoneme Group
True Positives	23	36	18
False Negatives	37	24	42
Sensitivity	38%	60%	30%
True Negatives	67	66	70
False Positives	6	7	3
Specificity	92%	90%	96%
+ Likelihood Ratio	4.66 *	6.26 *	7.30 *
- Likelihood Ratio	0.67	0.44	0.73

* Moderately strong effects (Dollaghan, 2007).

Discussion

The phonemes in the 2PL model did not overlap greatly with those used to calculate percentile scores on the GFTA-2. This is not surprising, as the 2PL phonemes were based on difficulty and discrimination, not "a controlled sample ... of the most frequently occurring consonant sounds in Standard American English" (Goldman & Fristoe, 2000, p. 7). Future screening measures of children's speech production skills may wish to consider difficult phonemes in difficult positions.

The predictive equations identified children with typical speech development with high accuracy and children with speech sound disorders with low accuracy. Although correctly producing the phonemes in the 3-, 5-, and 10-phoneme groups appears to have served as a proxy of the typically developing children's intact phonological systems, inaccurate productions of these same phonemes were unable to fully represent the variety of ways that errors can be present in children with speech sound disorders.

The primary next steps focus on similar analyses of larger data sets. This would lead to more complex modeling, such as a 3PL model and polytomous IRT models. Other areas for exploration include phonemic descriptions of errors and examinations of if and how IRT results may vary across age groups.

Contact information:
tbracke@bgsu.edu

The data reported here were collected in studies that were funded by NIH R03 DC005702 (Munson) and NIH R03 DC006545 (Storkel).

Applying Item Response Theory to the Development of a Screening Adaptation of the Goldman-Fristoe Test of Articulation-2

Tim Brackenbury, Michael Zickar, Bowling Green State University, Benjamin Munson, Holly Storkel, KU

Selected References

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., 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., Berger, M., Kremer, B., & Anteunis, L. (2015). Quantification of experienced hearing problems with item response theory. *American Journal of Audiology, 22*, 252-262.

del Toro, C. M., Bislick, L. P., Comer, M., Velozo, C., Romero, S., Gonzalex Rothi, L. J., & Kendall, D. L. (2011). *Journal of Speech, Language, and Hearing Research, 54*, 1089-1100.

Embretson, S. E. & Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Law, J., Boyle, J., Harris, F., Harkness, A. & Nye, C. (2007). 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., & Blesesc, 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.

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

Nelson, H. D., Nygern, 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.

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.

Contact information:
tbracke@bgsu.edu

Poster presented at the Symposium on Research in Child Language Disorders
June 18., 2016