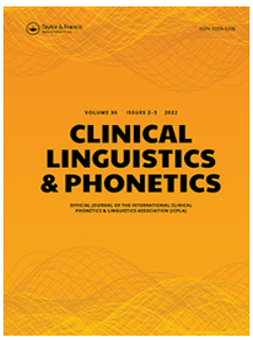

Clinical Linguistics & Phonetics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/iclp20

Automated phonological analysis and treatment target selection using AutoPATT

Philip Combiths, Ray Amberg, Gregory Hedlund, Yvan Rose & Jessica A. Barlow

To cite this article: Philip Combiths, Ray Amberg, Gregory Hedlund, Yvan Rose & Jessica A. Barlow (2022) Automated phonological analysis and treatment target selection using AutoPATT, Clinical Linguistics & Phonetics, 36:2-3, 203-218, DOI: 10.1080/02699206.2021.1896782

To link to this article: https://doi.org/10.1080/02699206.2021.1896782

Published online: 04 Jun 2021.

CLINICAL LINGUISTICS & PHONETICS
2022, VOL. 36, NOS. 2–3, 203–218
https://doi.org/10.1080/02699206.2021.1896782

Taylor & Francis
Taylor & Francis Group

Automated phonological analysis and treatment target selection using AutoPATT

Philip Combiths[a,b], Ray Amberg[c], Gregory Hedlund[d], Yvan Rose[d], and Jessica A. Barlow[a]

[a]School of Speech, Language, and Hearing Sciences, San Diego State University, San Diego, California, USA; [b]Department of Cognitive Science, University of California, San Diego, California, USA; [c]Alta Loma School District, Rancho Cucamonga, California, USA; [d]Department of Linguistics, Memorial University of Newfoundland, St. John's, Canada

ABSTRACT

Automated analyses of speech samples can offer improved accuracy and timesaving advantages that streamline clinical assessment for children with a suspected speech sound disorder. In this paper, we introduce AutoPATT, an automated tool for clinical analysis of speech samples. This free, open-source tool was developed as a plug-in for Phon and follows the procedures of the Phonological Analysis and Treatment Target Selection protocol, including extraction of a phonetic inventory, phonemic inventory with corresponding minimal pairs, and initial consonant cluster inventory. AutoPATT also provides suggestions for complex treatment targets using evidence-based guidelines. Automated analyses and target suggestions were compared to manual analyses of 25 speech samples from children with phonological disorder. Results indicate that AutoPATT inventory analyses are more accurate than manual analyses. However, treatment targets generated by AutoPATT should be viewed as suggestions and not used to substitute necessary clinical judgement in the target selection process.

ARTICLE HISTORY
Received 27 October 2020
Revised 17 February 2021
Accepted 23 February 2021

KEYWORDS
Phonology; phonological disorder; automated assessment

Introduction

Thorough phonological assessment is critical for identifying the presence, nature, and severity of speech sound disorders (SSDs), and for identifying appropriate treatment targets and goals. However, thorough assessment is time-consuming, and even the recommended 1–1.5 hours of direct assessment can be insufficient (Bleile, 2002; Miccio, 2002; Skahan et al., 2007). This is in addition to time spent post-assessment in analysis, determination of treatment goals, and paperwork, which is frequently reported to be more time-consuming than the assessment itself (Skahan et al., 2007). In total, most SLPs spend between 2 and 2.5 hours in the speech assessment and post-assessment process for children with suspected SSDs (McLeod & Baker, 2014; Skahan et al., 2007). As the global demand for SLPs serving children with SSDs continues to increase (American Speech-Language Hearing Association, 2016, 2020; Jesus et al., 2017; McAllister et al., 2013; Siewert et al., 2014), the efficiency of thorough diagnostic methods for suspected SSDs is an increasingly pressing concern.

CONTACT Philip Combiths pcombiths@sdsu.edu School of Speech, Language, and Hearing Sciences, San Diego State University, San Diego, CA 92182.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

In the context of time and resource restraints, short, standardized articulation tests, such as the Goldman-Fristoe Test of Articulation-3 (GFTA-3; Goldman & Fristoe, 2015), are often used to assess speech production, identify SSDs, and even determine treatment targets (Fabiano-Smith, 2019; McLeod & Baker, 2014; Skahan et al., 2007). However, because these types of standardized measures are limited in scope and heavily focused on relational analyses (i.e. comparing the child's productions to a correct target or a normative database), they fail to fully describe a child's speech as it is used in their day-to-day production. These measures do not provide a sample size with sufficient depth or breadth to conduct independent analyses, such as the establishment of phonetic or phonemic inventories (e.g., Barlow & Gierut, 2002; Eisenberg & Hitchcock, 2010; M.F. Elbert & Gierut, 1986; Stoel-Gammon, 1985), which describe a child's complete set of speech sound productions or determine their phonemic contrasts (Combiths et al., 2019; Eisenberg & Hitchcock, 2010; Fabiano-Smith, 2019; Macrae, 2017). Larger elicitation probes or connected speech samples collected through elicitation or spontaneous production are thus increasingly included in speech assessment as more thorough or naturalistic options (Bankson et al., 2017; Bernhardt & Holdgrafer, 2001; Macrae, 2017; Masterson et al., 2005; Miccio, 2002), but time is a concern when clinicians must transcribe and analyse these samples to extract useful diagnostic information from them.

Computerized tools for phonological analysis may offer at least a partial solution to the time investment required for a thorough speech assessment; however, very few SLPs report using these tools (McLeod & Baker, 2014; Skahan et al., 2007). Infrequent adoption of these technologies may be attributable to the limited scope, availability, or accessibility of tools that have been developed. Per Skahan et al. (2007), the most frequently used computerized phonological assessment tool is Hodson's Computerized Analysis of Phonological Patterns (2003). This commercially available software elicits a 50-word speech sample, automates a relational analysis of phonological error patterns (e.g., substitutions, cluster reductions), and provides error-pattern-based treatment target recommendations. Thus, this tool facilitates a rapid analysis of error patterns in English; however, it is not designed for independent analyses or for use with different types of speech samples. A similar program also exists for Malayalam, Computerized Assessment of Phonological Process in Malayalam (Sreedevi et al., 2013). Other commercially distributed tools, such as Logical International Phonetics Program (Oller & Delgado, 2000), Computerized Profiling (Long et al., 2006), and Computerized Articulation and Phonology Evaluation System (Masterson & Bernhardt, 2001) more flexibly facilitate transcription, analysis, and/or treatment target recommendation; however, these programs are no longer maintained and are either unavailable or incompatible with many modern devices.

Otherwise, only a few computerized phonological assessment tools are currently available and compatible with modern hardware. These include Programs to Examine Phonetic and Phonological Evaluation Records (PEPPER; Shriberg, 1990), maintained by the Weissman Center at the University of Wisconsin-Madison; Ferramentas para Análise Fonológica Automática [Automatic Phonological Analysis Tools] (APAT; Saraiva et al., 2017), maintained by the University of Aveiro, Portugal; and Phon (Rose & Hedlund, 2020; Rose & MacWhinney, 2014), which is part of TalkBank (MacWhinney, 2007) and maintained by Memorial University of Newfoundland. APAT is freely available (at http://acsa.web.ua.pt/) and completes analyses and produces results within Excel, which makes this tool readily accessible to users familiar with that software. Currently,

APAT is streamlined for speakers of European Portuguese using samples from the Teste Fonético-Fonológico–Avaliação da Linguagem Pré-Escolar [Phonological Testing–Pre-School Language Assessment] (Mendes et al., 2009) or the Teste de Articulação Verbal [Verbal Articulation Test] (Guimarães et al., 2014). Phon and PEPPER are stand-alone programs with graphical interfaces for transcription and analysis. Both are freely available (at https://www.phon.ca/ andhttps://phonology.waisman.wisc.edu/, respectively) and can accommodate a variety of speech sample types, including longer samples from independent probes or connected speech, to conduct clinically relevant analyses. Of these programs, Phon has been most recently updated. Because Phon is relatively accessible and actively maintained, its potential for improving the efficiency of speech assessments merits further examination.

To date, Phon has been most frequently used in research (Rose & Stoel-Gammon, 2015); however, it is also appropriate for clinical assessment and monitoring (Byun & Rose, 2016). Through a graphical user interface, Phon allows utterance segmentation (time alignment) as well as orthographic and phonetic transcription of connected speech or elicited samples of any length. Several of these steps can be automated or partially automated within Phon, which includes International Phonetic Alphabet (IPA) dictionaries and syllabification algorithms for multiple languages. These allow automated generation of target (model) transcriptions and phone-by-phone alignments between target and actual forms, all of which are automatically annotated for syllable-level information (e.g., syllable onsets or codas; syllable stress). Phon includes the capacity to conduct acoustic analysis through integration with Praat (Boersma & Weenink, 2020), and offers a number of clinically useful analyses, especially relational analyses, such as consonant accuracy/Percentage of Consonants Correct (PCC; Shriberg et al., 1997; Shriberg & Kwiatkowski, 1982) and phonological pattern analysis. Additionally, through a scripting language adapted for phonological queries, Phon permits customizable parsing of phonological data. Although certain independent inventory analyses that can provide a more complete description of a child's speech production are not currently integrated into Phon, these can be added with user-created scripts or plug-ins written in JavaScript or Groovy.

As described, Phon can be clinically useful given its ability to accommodate larger speech samples, partial automation of transcription, and built-in relational analyses; however, its current utility could be improved with the capacity to conduct additional independent analyses. Comprehensive independent analyses are often indicated as part of a thorough phonological assessment (e.g., Miccio, 2002; Skahan et al., 2007; Williams, 2015). For example, Phonological Analysis and Treatment Target (PATT) Selection procedures (Barlow et al., 2010) guide clinicians to conduct several independent analyses, including generating a phonetic inventory, an initial cluster inventory, and a phonemic inventory within a generative phonological framework (Chomsky & Halle, 1968). Together, these analyses provide a useful overview of the child's phonological system without relational comparisons to a correct model. From the results of these analyses, PATT procedures provide instructions for identifying gaps in a child's phonological knowledge and selecting relatively complex treatment targets. The recommendation of relatively complex targets is based on research which suggests that treatment targeting complex phonological structures results in greater system-wide phonological growth than targeting simpler structures (Elbert & McReynolds, 1979; Elbert et al., 1984; Flint & Costello Ingham, 2005; Gierut, 1990, 1991, 1998a, 1999; Gierut et al., 1987; Gierut & Morrisette, 2012; Gierut et al., 1996; Gierut &

Neumann, 1992; Pagliarin et al., 2009; Powell & Elbert, 1984; Powell et al., 1991; Sommers et al., 1967; Williams, 1991; cf. Rvachew & Bernhardt, 2010).

In order to supplement the clinical utility of Phon and provide more comprehensive independent analyses, we developed AutoPATT (available at https://github.com/rayamberg/AutoPATT) as a Groovy plug-in for Phon. Because PATT steps are procedural in nature (as described below), AutoPATT is able to replicate much of the manual process via automation. Given IPA transcription from a Phon session, AutoPATT automatically generates a phonetic inventory, a set of minimal pairs identifying phonemic contrasts, a phonemic inventory, and an initial cluster inventory. In keeping with PATT protocol, AutoPATT also generates a set of recommended treatment targets based on gaps in a child's phonological knowledge, as identified from the results of its inventory analyses.

Automated procedures for phonological analysis, such as those conducted by AutoPATT and other similar tools, could provide faster and more accurate speech assessment, although this has not been frequently studied. In one existing study, Saraiva et al. (2017) found that computerized APAT results were highly consistent with results derived manually from a standardized phonological assessment. Otherwise, there is a paucity of work in this area. Most automated phonological analyses have not been tested empirically, perhaps because the accuracy of automated procedures is taken for granted. Nevertheless, one cannot assume the accuracy of automated analyses, phonological or otherwise, because computerized processes can and do produce errored results.

Computational error is generally more systematic than human error (Hirschman & Mani, 2003; Strik & Cucchiarini, 2014), which tends to be more sporadic and unpredictable (McBride et al., 2014; Reason, 2000). When unexpected results arise with digital automation, these are usually the result of an error or oversight in the program's specified procedures, as a computer program is literal in its interpretation of instructions. Programs that are tested appropriately can avoid these systematic errors, allowing them to be used repeatedly while yielding results with consistently high levels of dependability. This is something we cannot expect from human operators, especially given the high degree of descriptive precision involved in the computation of many independent analyses, such as those completed by AutoPATT. Similarly, treatment target suggestions could be derived more systematically from automated algorithms, given appropriate and programmatic procedures. In sum, automated processes require testing and validation with realistic datasets to minimize potential systematic error, confirm intended results, and establish accuracy.

The current study

The necessity for identifying the accuracy of automated procedures and comparing them against manual procedures motivated the current study. To provide initial validation of AutoPATT analysis results, we compared computerized independent analyses and target selection with AutoPATT to those same procedures completed manually and identified the accuracy of these analyses using 25 speech samples from young children with phonological disorder. With this study, we seek to answer the following questions:

(1) Are automated phonetic, phonemic, and initial cluster inventories, as generated by AutoPATT, comparable to those same analyses conducted manually following PATT procedures?

(2) Does the accuracy of automated phonetic, phonemic, and initial cluster inventories, as generated by AutoPATT, differ from the accuracy of those same analyses conducted manually following PATT procedures?
(3) Are qualitative differences observable between AutoPATT target recommendations and targets generated manually following PATT procedures?

With this work, we contribute to the limited body of research investigating the accuracy of automated phonological analysis. Although treatment target selection is a component of both PATT and AutoPATT, subjective aspects of target selection and differences between manual and automated procedures make their accuracy difficult to quantify. Nevertheless, we observe target selection via both methods and compare them qualitatively.

Method

Participants and transcriptions

Participants in this study were 25 monolingual English-speaking children (age range = 3;1–6;7; mean age = 4;3) with functional phonological disorder (i.e. impairment in the production, acquisition, or representation of speech sounds with no known cause; Gierut, 1998b) from the Developmental Phonologies Archive of the Learnability Project[1] (Gierut, 2015b). Raw data were narrow phonetic transcriptions of each child's single-word productions from the Phonological Knowledge Probe (PKP; Gierut, 1985), collected prior to their participation in treatment. Reliability for 10% of consonant transcriptions was reported at 93% (Gierut, 2015a). The PKP samples 293 words (for wordlist, see Gierut, 2015c), with a minimum of five opportunities for each English phoneme, in each permissible word position. The PKP is also designed to elicit minimal pairs from which an individual's phonemic contrasts can be established. To permit analyses of these data with AutoPATT, transcriptions were converted from their archival format to a format compatible with Phon (for further description of this process, see Combiths et al., 2019)

Automated and manual data

From these transcriptions, two types of data were derived to compare agreement across manually generated analyses and automated analyses. For the manual analyses, research assistants in a phonology research laboratory were trained to manually complete PATT analysis procedures. Each research assistant demonstrated proficiency with these procedures using a sample dataset prior to contributing to the study. After this training, research assistants completed the PATT for each of 25 samples. PATT assessment procedures include generating:

[1]Archival data were retrieved from the Gierut/Learnability Project collection of the IUScholarWorks repository at https://scholarworks.iu.edu/dspace/handle/2022/20061 The archival data were original to the Learnability Project and supported by grants from the National Institutes of Health to Indiana University (DC00433, RR7031K, DC00076, DC001694; PI: Gierut). The views expressed herein do not represent those of the National Institutes of Health, Indiana University, or the Learnability Project. The author(s) assume(s) sole responsibility for any errors, modifications, misapplications, or misinterpretations that may have been introduced in extraction or use of the archival data.

(1) a phonetic inventory based on a two-time occurrence in the sample, with a corresponding list of English phones missing from the inventory
(2) a list of minimal pairs demonstrating phonemic contrasts
(3) a phonemic inventory derived from a two-time occurrence of minimal pairs, with a corresponding list of English phonemes missing from the inventory
(4) an inventory of word-initial consonant clusters based on a two-time occurrence in the sample, with a corresponding list of English clusters missing from the inventory

Although only these independent inventory analyses were evaluated quantitatively for the purposes of this study, PATT also includes a more involved complexity-based treatment target selection process (see Gierut, 2007; Morrisette et al., 2006; Storkel, 2018). In abbreviated form, this process includes:

(1) determining if any three-element consonant clusters (e.g. /spl-/) are appropriate targets based on their absence in a child's initial cluster inventory and the presence of components of the cluster (e.g./p/and/l/) in their phonemic inventories (Gierut & Champion, 2001)
(2) determining if any two-element consonant clusters (e.g. /fɪ-/) are appropriate targets, based on their absence in the initial cluster inventory, and their complexity relative to other English consonant clusters (Gierut, 1999)
(3) in the absence of potential cluster targets, determining a relatively complex singleton target (e.g. /θ/) based on absence from the phonetic inventory (e.g., Gierut et al., 1987), stimulability (e.g., Miccio et al., 1999), frequency, and age of acquisition (e.g., Gierut et al., 1996)

To generate the automated results, the aforementioned analyses and target selection steps were also completed using AutoPATT, which replicates these same procedures. Resultant inventories and sets of suggested targets were arranged such that each segment or cluster in an inventory or set of targets constituted an item for comparison purposes.

To determine accuracy of manual and automated inventories, the "correct" inventories were generated as follows. Each instance of disagreement between manual and automated analyses was reviewed by one of the authors. Referencing PATT procedures and the original raw data, the inclusion of a given segment or cluster in the phonetic, phonemic, or cluster inventory or set of suggested treatment targets was determined. A different author, blind to the initial designations, made accuracy determinations for 20% of the disagreements. When compared, reliability for these designations was 100%. In instances of agreement, the convergence of manual and automated results determined inclusion of that segment or cluster in the corresponding inventory.

Analyses

In order to compare automated analyses to manual analyses, several metrics of interrater reliability were calculated, including percent agreement, Cohen's kappa (Cohen, 1960), and Scott's pi (Scott, 1955). Cohen's kappa and Scott's pi are suitable measures for categorical data from two coders, accounting for the probability of chance agreement in the data (e.g. Mitani & Nelson, 2017). To determine the accuracy of each, these metrics were also calculated between automated and correct

Table 1. Interrater reliability for AutoPATT and manual analyses as percent agreement, Cohen's kappa, and Scott's pi.

Analysis	*n*	Auto-Manual % Agreement	*SE*	Cohen's Kappa	*SE*	Scott's Pi	*SE*
Phonetic Inventory	552	95.8%	0.009	−0.016	0.207	−0.021	0.209
Phonemic Inventory	398	89.2%	0.016	−0.014	0.146	−0.057	0.152
Cluster Inventory	108	75.9%	0.041	−0.105	0.189	−0.137	0.195
Total	1058	91.3%	0.009	−0.026	0.102	−0.045	0.104

inventories and between manual and correct inventories. Mixed effects logistic regression determined the ability of the automated results and manual results to predict correct outcomes, controlling for participant as a random factor.

Results

Agreement

Reliability between AutoPATT and manual analyses are displayed in Table 1 as percent agreement, Cohen's kappa, and Scott's pi values. Cohen's kappa and Scott's pi values near 0 are indicative of chance agreement (1 indicates perfect agreement). Percent agreement was highest for phonetic inventories (96%), followed by phonemic inventories (89%), and cluster inventories (76%). Despite these results, negative Cohen's kappa and Scott's pi values for all analyses suggest that agreement between AutoPATT and manual analyses are quite poor, given the relatively high probability of chance agreement in these data.

Accuracy

After considering the comparability of AutoPATT and manual analyses, we examined the relationship between results from AutoPATT and the results verified as correct according to PATT protocol, displayed in Table 2. Here, percent correct was high for all analyses: 100% or nearly 100% for phonetic and phonemic inventories, and 98% for cluster inventories. Cohen's kappa and Scott's pi indicated high agreement with correct results for phonetic, phonemic, and cluster inventories. Logistic regression indicated that, overall, AutoPATT results were a significant predictor of correct results, $z(24, 1033) = 2.75$, $p < .01$.

The relationship between results from manual analyses and the results verified as correct was examined in the same fashion, and these results are displayed in Table 3. Manual analyses were less accurate than AutoPATT analyses, with correct agreement at 96% for phonetic inventories, 89% for phonemic inventories, and 78% for cluster inventories. Cohen's kappa and Scott's pi indicated poor agreement for all analyses. Logistic regression indicated that, overall, manual results were not a significant predictor of correct results, $z(24, 1033) = 0.38$, $p = .71$.

Table 2. AutoPATT analysis reliability as percent correct, Cohen's kappa, and Scott's pi.

Analysis	*n*	AutoPATT % Correct	*SE*	Cohen's Kappa	*SE*	Scott's Pi	*SE*
Phonetic Inventory	552	99.8%	0.002	0.908	0.092	0.908	0.092
Phonemic Inventory	398	100.0%	0.000	1.000	0.000	1.000	0.000
Cluster Inventory	108	98.1%	0.013	0.865	0.094	0.865	0.095
Total	1058	99.7%	0.002	0.908	0.053	0.908	0.053

Table 3. Manual analysis reliability as percent correct, Cohen's kappa, and Scott's pi.

Analysis	*n*	Manual % Correct	*SE*	Cohen's Kappa	*SE*	Scott's Pi	*SE*
Phonetic Inventory	552	96.0%	0.008	−0.014	0.212	−0.020	0.213
Phonemic Inventory	398	89.2%	0.016	−0.014	0.146	−0.057	0.152
Cluster Inventory	108	77.8%	0.040	0.034	0.174	0.015	0.178
Total	1058	91.6%	0.009	0.017	0.100	0.050	0.087

Qualitative results

Quantitative analyses captured the overall relationship between AutoPATT and manual analysis results and provided an estimate of the accuracy of each; however, these did not provide insight into the sources of disagreement between manual and automated results or sources of error in either. For this we examined, qualitatively, the nature of discrepancies between AutoPATT and manual analysis results and their errors relative to correct results. These errors are displayed in Table 4. Furthermore, differences between AutoPATT and manual target selection were only examined qualitatively.

Most disagreements between AutoPATT and manual analyses (approximately 85%) were attributable to omission of a phone, phoneme, cluster, or treatment target from the relevant inventory or set of targets from the manual analysis. For phonetic and cluster inventories, these manual omissions were most common for non-ambient (i.e. not typically occurring in the target language) segments and clusters (e.g. [ʦ], [θw]) or segments and clusters with diacritic markers (e.g., [b̥], [dˢw]). Manual omissions from the phonemic inventory were frequently related to missing a second occurrence of minimal pairs for a given contrast or not identifying minimal pairs for non-ambient segments. Most of these omissions were classified as errors in the manual analysis.

Instances in which AutoPATT omitted an inventory item that was included in manual results were less common. These were most often attributable to differences in the interpretation of a phone or cluster for the purposes of inventory inclusion. For instance, [kj] occurred two or more times as a word-medial cluster in several participants' productions. On three occasions, research assistants included [kj] in the initial cluster inventory, based on these word-medial occurrences. However, PATT guidelines specify that inclusion in the cluster inventory should be based on a two-time occurrence of the cluster in word-initial position, making inclusion of [kj] in those cases an error. In a different example, [dð] occurred multiple times in the data. Based on its patterning in the samples, this was most likely meant to represent a dentalized affricate (i.e. [d͡ð]). Research assistants correctly interpreted these transcriptions as affricates, whereas the AutoPATT analysis erroneously interpreted these transcriptions as occurrences of clusters.

The number of manually selected treatment targets differed greatly across manual and automated procedures. This is primarily because most research assistants indicated only one treatment target, whereas AutoPATT provided a list of targets when multiple targets were appropriate. However, the manually selected target was always included in the set of potential targets identified by AutoPATT, with one exception. For one sample, the manually selected treatment target was /skw-/. However, AutoPATT missed this three-element cluster target because it erroneously considered an occurrence of a two-element cluster with a diacritic, [s̪t-], as an instance of a three-element cluster, eliminating potential three-element cluster targets. Identification of these discrepancies and errors provided useful information for later revisions of the AutoPATT algorithm, as discussed in the next section.

Table 4. Unique inventory errors.

Phonetic Inventory		*Phonemic Inventory*		*Cluster Inventory*	
Error Source	Errored Item	Error Source	Errored Item	Error Source	Errored Item
manual omission	d͡z	manual addition	ɾ	manual addition	ʔj
manual omission	d͡ʒ	manual omission	wˈ	manual omission	dᵊw
manual omission	nː	manual omission	d͡ʒ	manual omission	bᵊw
manual addition	t͡s̪	manual omission	b̥	manual omission	fw
manual omission	d̪	manual omission	p	manual omission	sw
manual omission	ð̥	manual omission	z	manual omission	bw
manual omission	d͡z̪	manual omission	tˀ	manual omission	dwˈ
manual omission	l	manual omission	m	manual addition	kj
manual omission	f	manual omission	h	manual omission	θw
manual omission	jᵊ	manual omission	ɹ	manual omission	ɡw
manual addition	d͡z̪	manual omission	k	manual omission	dw
manual omission	dᵊ	manual omission	w	manual addition	tw
manual addition	b̥	manual omission	ɡ	manual addition	dw
manual addition	ð	manual omission	b	manual addition	θn
manual addition	t͡ʃ	manual omission	t͡s̪	manual omission	ʃn
manual omission	d̥	manual addition	b̥	AutoPATT addition	dð
manual omission	nᵗ	manual omission	d̥	AutoPATT addition	tθ
manual omission	ɾ	manual omission	z̥		
manual omission	ɹᵗ	manual omission	ʃ		
manual omission	t̄	manual omission	ʔ		
manual omission	ʔ	manual omission	t͡s		
AutoPATT omission	dð	manual addition	v		
		manual omission	s̪		
		manual omission	s		
		manual omission	d͡z̪		
		manual omission	t͡ʃ		
		manual omission	ɾ		
		manual omission	j		
		manual omission	d͡z		
		manual omission	bˈ		
		manual omission	t̄		
		manual omission	tʰ		

Note: Only unique errors are displayed. Repeated errors are only listed once.

Discussion

In this study, we compared the results of manual and automated inventory analysis and treatment target selection, following PATT procedures (Barlow et al., 2010). We first discuss the quantitative and qualitative results for generation of inventories separately from the qualitative results for treatment target selection, as our analyses and findings differed in these areas, and they diverge in their relevant considerations.

Results indicate that automated generation of phonetic, phonemic, and cluster inventories using AutoPATT is not equivalent to these same inventories generated manually by undergraduate and graduate students with training in these procedures, at least when using narrowly transcribed speech samples from young children with phonological disorder. Low percent agreement between manual and automated inventory analyses were confirmed by near-zero Scott's pi and Cohen's kappa values. However, accuracy and qualitative error analyses revealed that disagreements were primarily attributable to human error in the manually generated analyses, most frequently omission of a phone, phoneme, or cluster that was accurately identified by AutoPATT. Specifically, AutoPATT-generated inventories were 98–100% accurate, whereas manually generated inventories were 78–96% accurate. Thus, AutoPATT may be a more accurate and consistent means of generating inventories for speech analysis than manual procedures.

In addition to generating inventory analyses, PATT and AutoPATT both include procedures for selecting complex treatment targets (e.g. Morrisette et al., 2006), based primarily on an individual's phonetic, phonemic, and initial cluster inventories. Research assistants frequently indicated only a single suggested target, and AutoPATT indicated a set of potential targets where applicable. Although automated procedures may be able to provide a short list of potentially appropriate targets, AutoPATT is ultimately unable to incorporate the myriad factors involved in determining a single target, based on independent inventory analyses alone. Furthermore, we identified a systematic error in the AutoPATT algorithm that led to misinterpretation of a two-element cluster with a diacritic [s̪t-] as a three-element cluster for the purposes of target selection – although this was addressed in subsequent revisions to the program, as described in the next section. From these observations, we conclude that AutoPATT may be able to streamline complex target selection by narrowing the pool of potential targets, but it is not a substitute for necessary clinical judgement in treatment target selection, and its suggestions must be reviewed against assessment results in case of unexpected errors.

Clinical implications

Automated phonological analyses, such as the generation of phonetic, phonemic, and cluster inventories, show promise as accurate means of describing an individual's phonological system. As shown here, automated inventories can be generated with less error than those created manually. Human error is a well-documented phenomenon in phonological analysis for research purposes (e.g., Shriberg & Lof, 1991); however, it is less frequently addressed in the clinical domain. For phonological analysis, relatively high proficiency with IPA notation is required, in addition to some knowledge of phonological theory. Even when clinicians are able to conduct these analyses, they are unlikely to have access to the time and resources available to research assistants in a phonology laboratory which permit them to work carefully and review their analyses for errors. Assuming accurate digital transcriptions, automated analyses also allow effective archival of an individual's speech production abilities, which can be referenced and analysed repeatedly. Because analysis procedures are applied identically across datasets, results can be compared over time with confidence that the analyses were conducted consistently. Consequently, clinicians might stand to benefit considerably from the greater accuracy and consistency of automated phonological analyses for speech assessment.

These automated analyses could be more efficient than manually completed analyses, although comparison of the time spent preparing and conducting manual and automated analyses still requires direct investigation. AutoPATT analyses require a speech sample to be transcribed in Phon. Transcriptions may be completed in Phon from a video or audio recording or by hand. If transcription is completed by hand, it requires data entry, either directly into Phon using its built-in IPA map or indirectly with another IPA typing tool. For a clinician or researcher experienced with Phon and digital IPA transcription, this transfer can be completed in 30 minutes for a sufficiently thorough sample, such as the 293-word samples in this study, but it could take an hour or more in other circumstances. Although digital transcription requires an initial time investment, phonological analysis software may offer a significant return on investment, as any number of relational and independent analyses can be completed and repeated in minutes. For comparison, trained students and research assistants typically spent 1–2 hours to complete manual PATT assessment and target selection procedures.

Although a pool of treatment target options, as generated by AutoPATT, can be a useful tool for clinicians seeking to identify relatively complex treatment targets for a child with phonological disorder, these suggestions are based on limited, one-dimensional inventory analyses. The onus of treatment target selection still lies on the clinician who may choose to consider these or other target options in the context of a complete assessment, which may include automated analysis of a speech sample, but should also include a variety of other assessment measures (Fabiano-Smith, 2019; Kamhi, 1992; McLeod & Baker, 2014; Miccio, 2002).

Limitations and future directions

Although AutoPATT was shown to be relatively accurate for inventory analyses in its current state, there remain areas for improvement. Research assistants following manual PATT procedures at times interpreted initial consonant clusters as exclusively word-initial and other times included syllable-initial (word-medial) clusters in that category. AutoPATT considered only word-initial clusters for the initial cluster inventory. Indeed, syllable-initial clusters may be appropriate to include in the initial cluster inventory, and both PATT and AutoPATT procedures could be updated accordingly. Also, at least one instance of systematic error in the AutoPATT target selection algorithm was identified, although this only impacted target selection for one participant, and the error has since been corrected.

Currently, AutoPATT does not consider an individual's stimulability for sounds absent from the phonetic inventory in the target selection process because this cannot be determined from a single-word sample. However, stimulability is a relevant consideration in target selection (e.g. Miccio et al., 1999). Similarly, substantial discrepancies between manual and automated treatment target selection highlight the critical role of clinical judgement in this process. Since completion of this study, AutoPATT has been revised to address the error in target selection described above, to display more detailed information explaining the characteristics of the sampled phonological system that resulted in the given set of suggested complex targets, and to clarify the utility of stimulability testing and other analysis tools in conjunction with clinical judgement for evidence-based assessment and treatment target selection.

Replication of this work would improve our understanding of the utility of these automated analysis tools. Extension of this work to other computerized assessment tools, to other populations, and with more heterogeneous samples would be especially beneficial, as the

current findings are only generalizable to AutoPATT analyses of samples from young monolingual English-speaking children with phonological disorder. Furthermore, PATT procedures exist for Spanish, and AutoPATT was developed for both English and Spanish; however, the accuracy of AutoPATT analyses in Spanish remains to be examined. As these automated analyses are revised, they will require additional validation. This iterative process will continue to improve the available repertoire of clinically appropriate tools for phonological analysis.

Conclusion

Automated speech analysis is an emerging area of academic and clinical interest, and it is increasingly considered a useful tool for clinical speech assessment. Clinicians may choose to expand their repertoire of phonological assessment tools to include automated analyses; however, the accuracy and validity of these tools should be taken into consideration, and automated results should be interpreted in the context of other conventional assessment tools and each client's unique set of personal circumstances and priorities.

Acknowledgments

We thank the members of the Phonological Typologies Lab at San Diego State University, especially Sarah Covington, Stephanie Nguyen, Brooke Hendricks, Taylor Kubo, Diana Carreon, and Teresa Roquet for their assistance with this project.

Disclosure statement

The fifth author is an author of the Phonological Analysis and Treatment Target (PATT) Selection protocol. The first, second, and fifth authors are authors of AutoPATT. The third and fourth authors are authors of Phon. These tools and software programs are distributed freely. The authors have no relevant financial disclosures.

Funding

This work was supported by the National Institutes of Health under Grants F31 DC017697, R21 DC01720, and R01 HD051698-11

ORCID

Philip Combiths http://orcid.org/0000-0003-3683-1650

References

American Speech-Language Hearing Association. (2016). *2016 schools survey report: SLP caseload characteristics*. http://www.asha.org/research/memberdata/schoolssurvey/

American Speech-Language Hearing Association. (2020). *Supply and demand resource list for speech-language pathologists*. https://www.asha.org/uploadedFiles/Supply-Demand-SLP.pdf

Bankson, N. W., Bernthal, J. E., & Peter Flipsen, J. (2017). Assessment: Data collection. In J. E. Bernthal, N. W. Bankson, & J. Peter Flipsen (Eds.), *Articulation and phonological disorders: Speech sound disorders in children* (pp. 150–175). Pearson.

Barlow, J. A., & Gierut, J. A. (2002). Minimal pair approaches to phonological remediation. *Seminars in Speech and Language*, *23*, 57–68. https://doi.org/10.1055/s-2002-24969
Barlow, J. A., Taps, J., & Storkel, H. L. (2010). Phonological Assessment and Treatment Target (PATT) Selection [Assessment protocol].
Bernhardt, B. H., & Holdgrafer, G. (2001). Beyond the basics I. *Language, Speech, and Hearing Services in Schools*, *32*(1), 18–27. https://doi.org/doi:10.1044/0161-1461(2001/002)
Bleile, K. M. (2002). Evaluating articulation and phonological disorders when the clock is running. *American Journal of Speech-Language Pathology*, *11*(3), 243–249. https://doi.org/10.1044/1058-0360(2002/026)
Boersma, P., & Weenink, D. (2020). *Praat: Doing Phonetics by Computer*. (Version 6.1) [Computer program].
Byun, T. M. A., & Rose, Y. (2016). Analyzing clinical phonological data using phon. *Seminars in Speech and Language*, *37*(2), 85–105. https://doi.org/10.1055/s-0036-1580741
Chomsky, N., & Halle, M. (1968). *The sound pattern of english*. Harper & Row.
Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104
Combiths, P. N., Barlow, J. A., & Sanchez, E. (2019). Quantifying phonological knowledge in children with phonological disorder. *Clinical Linguistics & Phonetics*, *33*(10–11), 885–898. https://doi.org/10.1080/02699206.2019.1584247
Eisenberg, S. L., & Hitchcock, E. R. (2010). Using standardized tests to inventory consonant and vowel production: A comparison of 11 tests of articulation and phonology. *Language, Speech, and Hearing Services in Schools*, *41*, 488–503. https://doi.org/10.1044/0161-1461(2009/08-0125)
Elbert, M., & McReynolds, L. V. (1979). Aspects of phonological acquisition during articulation training. *Journal of Speech and Hearing Disorders*, *44*, 459–471. https://doi.org/10.1044/jshd.4404.459
Elbert, M. F., Dinnsen, D. A., & Powell, T. W. (1984). On the prediction of phonologic generalization learning patterns. *Journal of Speech and Hearing Disorders*, *49*, 309–317. https://doi.org/10.1044/jshd.4903.309
Elbert, M. F., & Gierut, J. A. (1986). *Handbook of clinical phonology: Approaches to assessment and treatment*. Singular Group.
Fabiano-Smith, L. (2019). Standardized tests and the diagnosis of speech sound disorders. *Perspectives of the ASHA Special Interest Groups*, *4*(1), 58–66. https://doi.org/10.1044/2018_PERS-SIG1-2018-0018
Flint, C. B., & Costello Ingham, J. (2005). Pretreatment stimulability and percentage of consonants correct as predictors of across-phoneme generalization. *Contemporary Issues in Communication Science and Disorders*, *32*, 53–63. https://doi.org/10.1044/cicsd_32_S_53
Gierut, J. A. (1985). *On the relationship between phonological knowledge and generalization learning in misarticulating children* [Doctoral dissertation, Indiana University].
Gierut, J. A. (1990). Differential learning of phonological oppositions. *Journal of Speech and Hearing Research*, *33*(3), 540–549. https://doi.org/10.1044/jshr.3303.540
Gierut, J. A. (1991). Homonymy in phonological change. *Clinical Linguistics & Phonetics*, *5*(2), 119–137. https://doi.org/10.3109/02699209108985509
Gierut, J. A. (1998a). Natural domains of cyclicity in phonological acquisition. *Clinical Linguistics and Phonetics*, *12*(6), 481–499. https://doi.org/10.3109/02699209808985239
Gierut, J. A. (1998b). Treatment efficacy: Functional phonological disorders in children. *Journal of Speech, Language, and Hearing Research*, *41*(1), S85–S100. https://doi.org/10.1044/jslhr.4101.s85
Gierut, J. A. (1999). Syllable onsets: Clusters and adjuncts in acquisition. *Journal of Speech, Language, and Hearing Research*, *42*(3), 708–726. https://doi.org/10.1044/jslhr.4203.708
Gierut, J. A. (2007). Phonological complexity and language learnability. *American Journal of Speech-Language Pathology*, *16*, 6–17. https://doi.org/10.1044/1058-0360(2007/XXX)
Gierut, J. A. (2015a). Experimental designs and protocols. *Learnability Project Working Papers*. Indiana University. https://doi.org/10.5967/K8CC0XMD
Gierut, J. A. (2015b). *Learnability project*. Indiana University. https://scholarworks.iu.edu/dspace/handle/2022/20061
Gierut, J. A. (2015c). Phonological protocols. In *Learnability Project Working Papers* (1-50). Indiana University. https://doi.org/10.5967/K87P8W9V

Gierut, J. A., & Champion, A. H. (2001). Syllable onsets II: Three-element clusters in phonological treatment. *Journal of Speech, Language, and Hearing Research*, *44*(4), 886–904. https://doi.org/10.1044/1092-4388(2001/071)

Gierut, J. A., Elbert, M., & Dinnsen, D. A. (1987). A functional analysis of phonological knowledge and generalization learning in misarticulating children. *Journal of Speech and Hearing Research*, *30* (4), 462–479. https://doi.org/10.1044/jshr.3004.432

Gierut, J. A., & Morrisette, M. L. (2012). Age of word acquisition effects in treatment of children with phonological delays. *Applied Psycholinguistics*, *33*(1), 121–144. https://doi.org/doi:10.1017/S0142716411000294

Gierut, J. A., Morrisette, M. L., Hughes, M. T., & Rowland, S. (1996). Phonological treatment efficacy and developmental norms. *Language, Speech and Hearing Services in Schools*, *27*(3), 215–230. https://doi.org/0161-1461/96/2703-0215

Gierut, J. A., & Neumann, H. J. (1992). Teaching and learning/θ/: A non-confound. *Clinical Linguistics & Phonetics*, *6*(3), 191–200. https://doi.org/10.3109/02699209208985530

Goldman, R., & Fristoe, M. (2015). *Goldman-Fristoe test of articulation 3*. Pearson Clinical Assessment.

Guimarães, I., Birrento, C., Figueiredo, C., & Flores, C. (2014). *TAV–Teste de Articulação Verbal [Test of Verbal Articulation]* Oficina Didáctica.

Hirschman, L., & Mani, I. (2003). Evaluation. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 414–429). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199573691.013.014

Jesus, L. M. T., Martinez, J., Valente, A. R., & Costa, M. C. (2017). Speech and language therapy service delivery: Overcoming limited provision for children. *Public Health*, *151*, 39–50. https://doi.org/10.1016/j.puhe.2017.06.013

Kamhi, A. G. (1992). Phonological assessment and treatment: The need for a broad-based model of phonological disorders. *Language, Speech, and Hearing Services in Schools*, *23*(3), 261–268. https://doi.org/10.1044/0161-1461.2303.261

Long, S., Fey, M., & Channell, R. (2006). *Computerized Profiling*. (Version 9.7. 0) [Computer software] Case Western Reserve University.

Macrae, T. (2017). Stimulus characteristics of single-word tests of children's speech sound production. *Language, Speech, and Hearing Services in Schools*, *48*(4), 219–233. https://doi.org/10.1044/2017_LSHSS-16-0050

MacWhinney, B. (2007). The talkbank project. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora* (pp. 163–180). Springer.

Masterson, J., & Bernhardt, B. (2001). *Computerized articulation and phonology evaluation system (CAPES)*. (Version 4) [Computer program] The Psychological Corporation.

Masterson, J. J., Bernhardt, B. H., & Hofheinz, M. K. (2005, August 1). A comparison of single words and conversational speech in phonological evaluation. *American Journal of Speech-Language Pathology*, *14*(3), 229–241. https://doi.org/10.1044/1058-0360(2005/023)

McAllister, L., Wylie, K., Davidson, B., & Marshall, J. (2013). The world report on disability: An impetus to reconceptualize services for people with communication disability. *International Journal of Speech-language Pathology*, *15*(1), 118–126. https://doi.org/10.3109/17549507.2012.757804

McBride, S. E., Rogers, W. A., & Fisk, A. D. (2014). Understanding human management of automation errors. *Theoretical Issues in Ergonomics Science*, *15*(6), 545–577. https://doi.org/10.1080/1463922X.2013.817625

McLeod, S., & Baker, E. (2014). Speech-language pathologists' practices regarding assessment, analysis, target selection, intervention, and service delivery for children with speech sound disorders. *Clinical Linguistics & Phonetics*, *28*(7–8), 508–531. https://doi.org/10.3109/02699206.2014.926994

Mendes, A., Afonso, E., Lousada, M., & Andrade, F. (2009). *Teste Fonetico–Fonológico-Avaliação da Linguagem Pré-Escolar (TFF-ALPE) [phonological testing–pre-school language assessment]*. Designeed.

Miccio, A. W. (2002). Clinical problem solving: Assessment of phonological disorders. *American Journal of Speech-Language Pathology*, *11*, 221–229. https://doi.org/10.1044/1058-0360(2002/023)

Miccio, A. W., Elbert, M., & Forrest, K. (1999). The relationship between stimulability and phonological acquisition in children with normally developing and disordered phonologies. *American Journal of Speech-Language Pathology*, *8*(4), 347–363. https://doi.org/10.1044/1058-0360.0804.347

Mitani, A. A., & Nelson, K. P. (2017). Modeling agreement between binary classifications of multiple raters in R and SAS. *Journal of Modern Applied Statistical Methods*, *16*(2), 277–309. https://doi.org/10.22237/jmasm/1509495300

Morrisette, M. L., Farris, A. W., & Gierut, J. A. (2006). Applications of learnability theory to clinical phonology. *Advances in Speech-Language Pathology*, *8*, 207–219. https://doi.org/10.1080/14417040600823284

Oller, K., & Delgado, R. (2000). *Logical International Phonetics Program* (Version 2) [Computer software] Intelligent Hearing Systems.

Pagliarin, K. C., Mota, H. B., & Keske-Soares, M. (2009). Análise da eficácia terapêutica em três modelos fonológicos de abordagem contrastiva [Therapeutic efficacy analysis of three contrastive approach phonological models]. *Pro-fono: Revista De Atualizacao Cientifica*, *21*(4), 297–302. https://doi.org/10.1590/S0104-56872009000400006

Powell, T. W., & Elbert, M. (1984). Generalization following the remediation of early- and later-developing consonant clusters. *Journal of Speech and Hearing Disorders*, *49*(2), 211. https://doi.org/10.1044/jshd.4902.218

Powell, T. W., Elbert, M., & Dinnsen, D. A. (1991). Stimulability as a factor in the phonological generalization of misarticulating preschool children. *Journal of Speech and Hearing Research*, *34*, 1318–1328. https://doi.org/10.1044/jshr.3406.1318

Reason, J. (2000). Human error. *Western Journal of Medicine*, *172*(6), 393–396. https://doi.org/10.1136/ewjm.172.6.393

Rose, Y., & Hedlund, G. (2020). *Phon*. (Version 3.1) https://www.phon.ca/

Rose, Y., & MacWhinney, B. (2014). The PhonBank project: Data and software-assisted methods for the study of phonology and phonological development. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 380–401). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571932.013.023

Rose, Y., & Stoel-Gammon, C. (2015). Using phonbank and phon in studies of phonological development and disorders. *Clinical Linguistics & Phonetics*, *29*(8–10), 686–700. https://doi.org/10.3109/02699206.2015.1041609

Rvachew, S., & Bernhardt, B. M. (2010, February 1). Clinical implications of dynamic systems theory for phonological development. *American Journal of Speech Language Pathology*, *19*, 34–50. https://doi.org/10.1044/1058-0360(2009/08-0047)

Saraiva, D., Lousada, M., Hall, A., & Jesus, L. M. T. (2017). Paediatric automatic phonological analysis tools (APAT). *Logopedics Phoniatrics Vocology*, *42*(4), 153–159. https://doi.org/10.1080/14015439.2016.1237544

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321–325. https://doi.org/10.1086/266577

Shriberg, L., Austin, D., Lewis, B. A., McSweeny, J., & Wilson, D. (1997). The Percentage of Consonants Correct (PCC) metric: Extensions and reliability data. *Journal of Speech, Language, and Hearing Research*, *40*, 708–722. https://doi.org/10.1044/jslhr.4004.708

Shriberg, L. D. (1990). *PEPPER: Programs to examine phonetic and phonological evaluation records*. [Computer program] Lawrence Erlbaum Associates.

Shriberg, L. D., & Kwiatkowski, J. (1982). Phonological disorders III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders*, *47*, 256–270. https://doi.org/10.1044/jshd.4703.256

Shriberg, L. D., & Lof, G. L. (1991). Reliability studies in broad and narrow transcription. *Clinical Linguistics & Phonetics*, *5*, 225–279. https://doi.org/10.3109/02699209108986113

Siewert, R., Boyce, S., Seselja, Z., Bilyk, C., Brown, C., Peris, N., Reynolds, L., Smith, D., Thorp, H. L., Urquhart, A., & Moore, C. (2014). The demand for speech pathology services in Australia. In *Prevalence of different types of speech, language and communication disorders and speech pathology services in australia [senate report]* (pp. 33–43). Commonwealth of Australia.

Skahan, S. M., Watson, M., & Lof, G. L. (2007). Speech-language pathologists' assessment practices for children with suspected speech sound disorders: Results of a national survey. *American Journal of Speech-Language Pathology*, *16*(3), 246–259. https://doi.org/10.1044/1058-0360(2007/029)

Sommers, R. K., Leiss, R. H., Delp, M. A., Gerber, A. J., Fundrella, D., Smith, R. M., Revucky, M. V., Ellis, D., & Haley, V. A. (1967). Factors related to the effectiveness of articulation therapy for kindergarten, first, and second grade children. *Journal of Speech and Hearing Research*, *10*, 428–437. https://doi.org/10.1044/jshr.1003.428

Sreedevi, N., John, M., & Chandran, S. (2013). Computerized Assessment of Phonological Processes in Malayalam (CAPP-M).

Stoel-Gammon, C. (1985). Phonetic inventories, 15-24 months: A longitudinal study. *Journal of Speech and Hearing Research*, *28*, 505–512. https://doi.org/10.1044/jshr.2804.505

Storkel, H. L. (2018). The complexity approach to phonological treatment: How to select treatment targets. *Language, Speech, and Hearing Services in Schools*, *49*(3), 1–19. https://doi.org/10.1044/2017_LSHSS-17-0082

Strik, H., & Cucchiarini, C. (2014). On automatic phonological transcription of speech corpora. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 89–109). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571932.013.001

Williams, A. L. (1991). Generalization patterns associated with training least phonological knowledge. *Journal of Speech and Hearing Research*, *34*, 722–733. https://doi.org/10.1044/jshr.3404.733

Williams, A. L. (2015). Assessment and intervention from a systemic perspective. In C. Bowen (Ed.), *Children's speech sound disorders* (2nd ed., pp. 199–203). Wiley-Blackwell.