JSLHR

Research Article

School-Aged Children's Phonological Accuracy in Multisyllabic Words on a Whole-Word Metric

Glenda K. Mason[a]

Purpose: The purpose of this study is to examine differences in phonological accuracy in multisyllabic words (MSWs) on a whole-word metric, longitudinally and cross-sectionally, for elementary school–aged children with typical development (TD) and with history of protracted phonological development (PPD).
Method: Three mismatch subtotals, Lexical influence, Word Structure, and segmental Features (forming a Whole Word total), were evaluated in 3 multivariate analyses: (a) a longitudinal comparison (n = 22), at age 5 and 8 years; (b) a cross-sectional comparison of 8- to 10-year-olds (n = 12 per group) with TD and with history of PPD; and (c) a comparison of the group with history of PPD (n = 12) with a larger 5-year-old group (n = 62).
Results: Significant *effect sizes* (η_p^2) found for mismatch totals were as follows: (a) moderate (Lexical, Structure) and large (Features) between ages 5 and 8 to 10 years, mismatch frequency decreasing developmentally, and (b) large between 8- to 10-year-olds with TD and with history of PPD (Structure, Features; minimal lexical influences), in favor of participants with TD. Mismatch frequencies were equivalent for 8- to 10-year-olds with history of PPD and 5-year-olds with TD. Classification accuracy in original subgroupings was 100% and 91% for 8- to 10-year-olds with TD and with history of PPD, respectively, and 86% for 5-year-olds with TD.
Conclusion: Phonological accuracy in MSW production was differentiated for elementary school–aged children with TD and PPD, using a whole-word metric. To assist with the identification of children with ongoing PPD, the metric has the ability to detect weaknesses and track progress in global MSW phonological production.

Recent studies suggest that multisyllabic words (MSWs) are important contexts for evaluation of a child's speech production, particularly beyond age 5 years (e.g., James, Ferguson, & Butcher, 2016; Mason & Bernhardt, 2014). The overarching aim of the current study, therefore, was to add to the limited data about phonological accuracy in MSWs[1] (i.e., words of two or more syllables), for school-aged children with typical phonology and with history of protracted phonological development[1] (PPD).

Development of Phonological Processes in MSW Accuracy

Researchers have described phonological development in MSW production in terms of the frequencies of phonological patterns/processes/errors. For instance, James (2006) outlined evidence concerning developmental patterns and growth in accuracy for 264 typically developing Australian children between ages 4 and 7 years. James observed that children under the age of 7 years showed mastery of word length and phoneme sequences but demonstrated stress pattern errors that subsequently resolved after age 7 years. James (2006) proposed a five-stage model of MSW acquisition, structured on observed phonological patterns, that is, stress and sequence alterations, and component deletion and addition. In the model, error pattern frequencies decreased as children got older; however, there was overlap of some patterns across stages. Furthermore, James (2006) and Kehoe (2001) noted irregular decreases in phonological patterns in MSWs. For example, sequence alterations apparently peaked between ages 4 and 5 years and then decreased, only to peak again at the age of 6 years. These patterns remained constant thereafter until 7 years of age, the oldest age group studied (James, 2006). The observed overlaps and irregularities highlighted developmental trade-offs between accuracy of word prosody and phonemes.

In another study, James's (2006) notion of stages was also invoked for describing atypical phonological

[a]School of Audiology & Speech Sciences, The University of British Columbia, Vancouver, Canada
Correspondence to Glenda Mason: gkellmas@audiospeech.ubc.ca

Editor-in-Chief: Julie Liss
Editor: Maria Grigos
Received April 13, 2017
Revision received December 4, 2017
Accepted July 20, 2018
https://doi.org/10.1044/2018_JSLHR-S-17-0137

[1] *PPD* (B. H. Bernhardt & Stemberger, 1998; Dubasik & Ingram, 2013) is used to encompass labels such as *Phonological/Speech Sound/(In) Consistent Speech/(Developmental) Speech + Disorder/Impairment/Delay/Errors*.

development of words of three or more syllables. That is, Masso, McLeod, Baker, and McCormack (2016) proposed a Framework of Polysyllable Maturity for ninety-three 4- to 5-year-olds with speech sound disorder. Masso et al. identified seven main phonological error categories (with 32 subcategories) from samples of 30 broadly transcribed polysyllabic words. Phonemic substitution, the most common error across the sample, was excluded from the model because it did not further distinguish stages of maturity. Instead, a five-level framework was structured on the next three most frequent patterns, that is, incorrect number of syllables and/or consonants, consonant or vowel deletion, and timing/stress inaccuracy. Just as James (2006) had observed, Masso et al. (2016) noted overlap of phonological patterns across levels of the framework. Nevertheless, participants were assigned to a level of polysyllable maturity on the basis of demonstrating each of the level's relevant error categories in 40% of words sampled. Because the resultant assignments produced a distribution of participants across all five levels, Masso et al. concluded that their framework was valid.

Theory and MSW Production

With respect to a valid model of children's phonological development in MSWs, two distinct theoretical views are relevant. The first view, as mentioned, is acquisition across stages comprising phonological patterns with decreasing frequency (James, 2006; Masso et al., 2016). The second view is a continuum of acquisition of interacting levels of hierarchically organized phonological structure (B. H. Bernhardt & Stemberger, 1998; Rvachew & Brosseau-Lapré, 2012). The next section elaborates on these theories and related issues.

Relative to the stages' view of phonological development in MSWs, James (2006) and Masso et al. (2016) suggested that phonological patterns reflect the accuracy of children's underlying representations, which could, in turn, be understood with respect to the accuracy of unique categories of phonological components. In the production of components, combinations of error patterns could be considered representative of developmental stages (James, 2006) or levels of maturity (Masso et al., 2016). However, in such modular views, the overlap of patterns/error categories across stages/levels creates ambiguity because stages/levels are not mutually exclusive. Furthermore, it is difficult to explain multiple error patterns that may occur within a word's production. For example, if a weakly stressed consonant–vowel–consonant syllable deletes from a word, stress is altered, and in addition, the segmental deletions alter phonotactics. The relative importance of patterns would require clarification, both those isolated in a particular stage and those with overlap. An additional issue specific to the Framework of Polysyllable Maturity (Masso et al., 2016) is the masking effects that can result from excluding error categories with lower or higher frequencies than those that are included.

To account for the likelihood of interactions and trade-offs among all MSW components, all observed patterns should be considered. The notion of interactions among word components can also account for multiple error categories within a single word production or across stages/levels. Discrete error categories are challenging to determine because all of the many components of an MSW are relevant to its output. Discreteness also implies instability of phonological representations when apparent trade-offs occur, for instance, when stress accuracy is achieved at the expense of seemingly established features, as illustrated below:

1. *balloon* /bəˈlun/ (a) [blun] (b) [bəˈwun]
   (a) off-target (disyllable) weakStrong stress, but on-target /l/.
   (b) on-target (disyllable) weakStrong stress, but off-target /l/, realized as /w/.
2. *animal* /ˈænəml/ (a) [ˈænml] (b) [ˈæməml]
   (a) off-target (three-syllable) Strongweakweak stress, but on-target /n/.
   (b) on-target (three-syllable) Strongweakweak stress, but off-target /n/, realized as /m/.

For explaining the complex dimensions of MSW phonological output and the changes and variability across the course of phonological development, a continuum view that considers a combination of linguistic models of nonlinear phonology and models of parallel interactive language processing seems better suited. Progression along the continuum depends on the pervasiveness and persistence of patterns of constraint on the phonological components. Phonological patterns observed in typical phonological development and, in addition, atypical patterns observed in PPD can be explained in terms of the multiple interacting levels of a hierarchically organized phonological system (B. H. Bernhardt & Stemberger, 1998). For example, from a linear standpoint, segments in words such as *hippopotamus* /ˌhɪpəˈpʰɑːrəmɪs/ and *cash register* /ˈkʰæːʃˌɹɛdʒɪstə-/ seem dissimilar in terms of number, manner, and early or late developmental acquisition (Smit, Hand, Freilinger, Bernthal, & Bird, 1990). A nonlinear analysis, however, highlights more of the similarities between higher levels of phonological chunks. For instance, both words contain two groups of syllables, each of which carries stress (i.e., a foot,[2] carrying either primary or secondary stress) and a StrongWeakweak stress pattern is attributed to the word-final foot of both words, that is, /ˌɹɛdʒɪstə-/ and /ˈpʰɑːrəmɪs/. The word-initial syllables, /ɹɛ/ and /pʰɑː/, also share onset-nucleus consonant–vowel structure, and in theory, the rime /ɪs/, may be common to both words. Accordingly, even if an MSW such as *hippopotamus* is composed of earlier acquired segments, insufficient hierarchical organization at higher levels of a phonological system might contribute to off-target output.

While nonlinear frameworks define the linguistic relationships of phonological forms, parallel interactive models of language processing (e.g., B. M. Bernhardt, Stemberger, & Charest, 2010; Wheeler & Touretzky, 1997)

[2]Defined as a group of two or more syllables in which one syllable carries stress (B. H. Bernhardt & Stemberger, 1998).

further delineate the relationships between parts of words during speech production. That is, during production of a lexical item, dynamic feed forward–feed backward interactions occur among semantic-based representations and hierarchically organized linguistic components, with varying amounts of neurological activation. One or more competing components could be active at a given point in processing, with underactivation provoking errors in selection. The relationships between segments and higher levels (e.g., syllable, rime, and nucleus) could be one to many or many to one, leaving the possibility of multiple errors in a single word production. Instability of phonological representations and trade-offs is possible because static representations are not stored in long-term memory; instead, a phonological representation is constructed each time a word is spoken (B. M. Bernhardt et al., 2010; Presson & MacWhinney, 2011). Adopting a view of parallel interactive language processing, therefore, precludes analysis of MSW production in terms of discrete components or error categories.

The Current Study

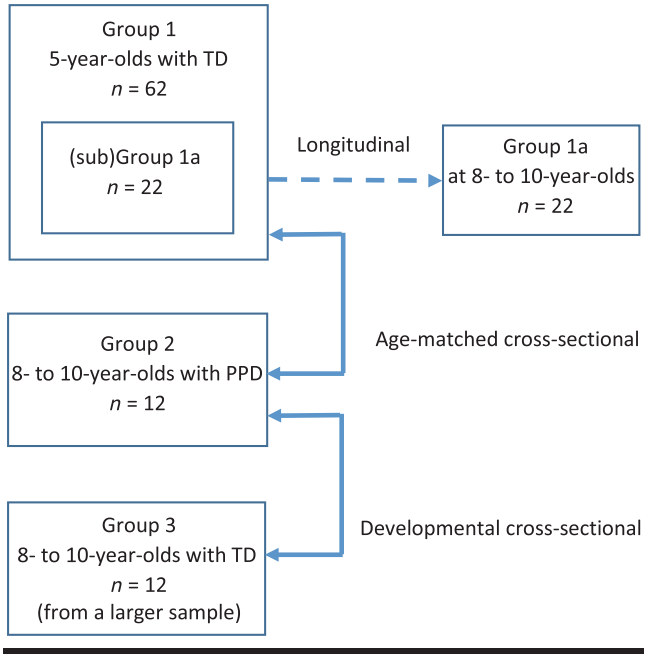
The goal of the current study was to contribute further information to the sparse research on phonological acquisition in MSWs by examining the phonological accuracy of elementary school children with typical development (TD) and with PPD. The study differs from the few existing studies of phonological production in MSWs in three ways: the focus on older children, the included word lengths, and the methodology for describing phonological inaccuracy. First, although none of the prior studies had evaluated children beyond the age of 7 years, the current study included 8- to 10-year-olds. Second, by definition, disyllables have been excluded from the evaluation of polysyllabic words[3] (e.g., James et al., 2016; Masso et al., 2016); however, disyllables have also been vulnerable to phonological inaccuracy (James, 2006) and thus were included for the present analysis.

Third, phonological inaccuracies in MSWs have traditionally been described in terms of discrete phonological processes. In contrast, for the current study, a whole-word metric (Mason, Bérubé, Bernhardt, & Stemberger, 2015) was applied. The metric was designed to integrate theories of language processing and hierarchical (nonlinear) phonological structure to provide a total for cumulative lexical selection and phonological mismatches (errors) across words. The mismatch total was summed from three mismatch subtotals for (a) lexical selection influences, (b) word structure components, and (c) segmental features. Lexical selection influences were tallied where mismatched syllables were members of (usually) monosyllabic word neighborhoods. When a lexical mismatch was tallied, no further phonological mismatch tallies were assigned unless word structure was affected. Word structure components comprised stress and syllable numbers and consonant–vowel shape. Segmental features included manner, place and laryngeal[3] for consonants, and height, backness, tenseness, nasalization, and roundness for vowels. Nonallophonic and nonassimilatory diacritics were also tallied, such that all inaccuracies were accounted for, but no consonant or vowel feature category was tallied more than once.

The aims of the study were addressed in three comparisons, one longitudinal and two that were cross-sectional. These analyses utilized data from three groups of children (see Figure 1). First, for the longitudinal analysis, the question was whether children with TD would demonstrate significant growth in phonological accuracy in MSWs during elementary school and whether frequencies would be more apparent in terms of total mismatches or for any subtotal. The aim was to document the predicted developmental progression for a subgroup of 5-year-olds with TD who were reevaluated at 8 years of age (Group 1a). Age was expected to affect the frequencies of the mismatch total and all three subtotals, with significantly higher frequencies at 5 than at 8 to 10 years of age (James, 2006; Kehoe, 2001). Because of the wide age range between groups, significant and large differences were generally predicted. For the lexical subtotal, however, moderate differences were expected because some MSWs might be less familiar to 5-year-olds.

The two cross-sectional analyses concerned phonological accuracy in MSWs for children with history of PPD, addressing two questions: (a) whether total mismatches or any subtotal would differ significantly from children of the same age, and (b) whether differences would be less in

Figure 1. Participant groups and comparisons. TD = typical phonological development; PPD = protracted phonological development.



[3]Defined as features that involve control of the larynx, including vocal cord vibration, spreading, and constriction, often referred to as voice (B. H. Bernhardt & Stemberger, 1998).

comparison with developmentally younger children. The aims were to demonstrate differences between 8- to 10-year-olds with history of PPD (Group 2) and two groups with TD: one, developmentally younger 5-year-olds (Group 1), and the other, age matched (Group 3). Whether or not 8- to 10-year-olds had history of PPD, the influence of word familiarity presumably would be less important (Storkel & Morrisette, 2002), such that lexical mismatch differences would be minimal between Groups 2 and 3; however, moderate differences were expected between Groups 2 and 1. Because the children in Group 2 were several years older, compared with Group 1, lower total mismatches and fewer word structure and feature mismatches were predicted but with small differences. Compared with Group 3 (age matched), Group 2 (history of PPD) was expected to have higher total mismatches and more word structure and feature mismatches with moderate differences predicted.

There were also two secondary aims of the study. The first of these was to add evidence to the validity of an MSW metric (Mason et al., 2015) that quantifies a wide range of prosodic and segmental components (Ingram, 2002; Kirk & Vigeland, 2015). The other aim was to contribute to the verification of the theoretical underpinnings of the metric, that is, the combination of theories of linguistic structure and language processing.

Method

Participants

All participants in the study were recruited from a small urban/rural school district in British Columbia, Canada, with an approximate enrollment of 6,500 students. The samples came primarily from middle-income homes and, overall, were representative of the school district's annual report of family income demographics. From responses provided in parent/caregiver questionnaires, the children were designated as typically developing based on unremarkable birth, medical, and developmental histories, including no enrollment in intervention for language or pragmatic communication. Children with typical speech development also had no reported history of speech therapy. By verbal report of the classroom teachers, the children with TD were meeting the expected learning outcomes and therefore not receiving individual education programs or other supplemental instruction. All participants, with TD and PPD, had passed a hearing screening from 500 to 4000 Hz bilaterally, usually the school district kindergarten screening. Children without a documented hearing screening were screened for hearing impairment by the school-based speech-language pathologist (SLP), following the procedural protocol of the district kindergarten screening program.

In order to sample children with TD, classroom teachers were asked to distribute an information letter about the study, as well as consent and assent forms, to parents of all eligible students. Between September and December 2007, 64 monolingual Canadian English–speaking 5-year-olds (Group 1) were recruited. As a result of lost data, two girls were subsequently removed from the sample, such that the remaining 62 children comprised 18% of the total kindergarten enrollment. This constituted the group of younger children (Group 1) to which the sample with PPD (Group 2) was compared. From September to December 2010, a sample of 8- to 10-year-olds with TD was also recruited, with a random sample of 64 children meeting the eligibility requirements (some for future study). For the longitudinal analysis, a subgroup of these children (Group 1a) who had participated in the 5-year-old sample were followed. For the cross-sectional comparison of 8- to 10-year-olds with history of PPD and with TD, three criteria were used to match the participants: (a) gender; (b) age, matched within 6 months; and (c) evaluation time (spring or fall). The resultant TD group (Group 3) did not overlap with the children in the longitudinal sample. Therefore, of the 64 eligible 8- to 10-year-olds, 30 were not selected to participate, either because they had not participated when 5 years old or because they could not be matched to a peer in the group with PPD.

In order to sample children with history of PPD (Group 2), between September and December 2013, all SLPs in the district were asked to distribute study information and consent forms to caregivers of children on their caseloads. The children were enrolled in kindergarten to Grade 5, the final elementary school year. The school-based SLPs, with access to the schools' confidential files, identified assessment documentation demonstrating that the children had history of PPD (as opposed to suspected *childhood apraxia of speech* or speech sound disorders of articulatory origin), whether or not considered resolved at the time of the study. All of the children had received speech therapy in the past. Either within the year previous or subsequent to kindergarten enrollment, a Canadian-registered SLP had evaluated the children's phonology. Identification of PPD was in terms of standardized percent consonants correct (PCC) scores at least 1 *SD* below the mean, for example, Structured Photographic Articulation Test II Featuring Dudsberry (Dawson & Tattersall, 2001) and/or a criterion-referenced assessment showing unexpected phonological processes for age, for example, Hodson Assessment of Phonological Patterns–Third Edition (Hodson, 2004). For language comprehension and expression, composite scores in both domains were within the average range on a standardized test battery, mean (*SD*) = 100 (15), for example, Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4; Semel, Wiig, & Secord, 2003) or Test of Language Development–Primary: Third Edition (Newcomer & Hammill, 1997). The sampling procedures resulted in a convenience sample of 21 children with history of PPD, with the youngest seven enrolled in kindergarten to Grade 2 (age range: 6;1 [years;months] to 7;4). These younger students were excluded from the study because the age range of the remainder (7;8 to 10;0) better approximated that of the 8- to 10-year-olds with TD. Of the remaining 14 children, two additional children were excluded on the basis of remarkable birth history (i.e., prenatal, perinatal, and/or postnatal complications), such that 12 were included in the final group with PPD.

For the three comparisons, participants in Group 1a were 22 children with TD (11 boys, 11 girls) who were evaluated at age 5 years (M = 5.6, SD = 3.08 months; range: 5;3–6;0) and followed up between 8 and 10 years of age (M = 9.44, SD = 6.35 months; range: 8;7–10;3). Next, for the two cross-sectional comparisons, Group 2 comprised twelve 8- to 10-year-olds (7 boys, 5 girls) with history of PPD (M = 9.38, SD = 9.64 months; range: 8;5–10;10) who were compared with different groups of children with TD: (a) Group 1, sixty-two 5-year-olds (28 boys, 34 girls), (M = 5.64, SD = 3.08 months; range: 5;3–6;0) and (b) Group 3, 12 matched typical-speaking 8- to 10-year-olds (7 boys, 5 girls), (M = 9.35, SD = 11.72 months; range: 7;8–10;10). The sample for the current study therefore comprised the sixty-two 5-year-olds (including 22 who participated in the longitudinal comparison) and the twenty-four 8- to 10-year-olds with TD and with PPD, for a total of 86 different children.

Procedure

Within 8 months of sample selection for a given comparison, testing occurred over two 1-hr sessions, evaluating the children at their schools on a wide variety of language and literacy-related measures (the latter for future study). Tasks requiring a verbal response were audio-recorded for scoring verification. The principle investigator, a Canadian-registered SLP, collected the data for all children with typical phonological development. Prior to data collection for the children with history of PPD, the principle investigator conducted two training sessions for two graduate SLP student research assistants. Subsequently, for each research assistant, procedural fidelity was observed in two evaluation simulations with school-aged nonparticipants with TD and for the authentic evaluations of two study participants with history of PPD. The principle investigator also verified the accuracy of scoring for all non–speech production tasks.

Consistent with customary pediatric clinical practice, task presentation order was balanced in terms of comprehension and production demands. In order to demonstrate that the comparison groups had average and equivalent standard scores for language-related skills, two tasks were administered: (a) vocabulary comprehension (Peabody Picture Vocabulary Test–Fourth Edition; Dunn & Dunn, 2007), mean (SD) = 100 (15), and (b) working memory (Number Repetition subtests of the CELF-4; Semel et al., 2003), mean (SD) = 7 (3). Regarding the CELF-4 (Semel et al., 2003), to challenge working memory, all 8- to 10-year-olds were administered *Number Repetition Backward*, but to avoid ceiling effects, the 5-year-olds were tested on *Number Repetition Forward*. In order to retain sample size, for below average standard scores on either task, a related measure was checked for average range performance. Vocabulary standard scores for both age groups were considered on the Expressive Vocabulary Test–Second Edition (Williams, 2007), mean (SD) = 100 (15). Concerning working memory, for the 5-year-olds, standard scores were judged on the Recalling Sentences subtest of the Clinical Evaluation of Language Fundamentals Preschool–Second Edition (Wiig, Secord, & Semel, 2004), mean (SD) = 7 (3). For the 8- to 10-year-olds with TD and PPD, working memory was considered in terms of standard scores on the Segmenting and Elision subtests of the Comprehensive Test of Phonological Processing (Wagner, Torgesen, & Rashotte, 1999), mean (SD) = 7 (3).

Phonological samples of MSWs were collected using the (Phonemic) Profile and Individualized Phonemic Evaluation–Level 4 (IPE4) of the Computerized Articulation and Phonology Evaluation System (CAPES; Masterson & Bernhardt, 2001). The Profile and IPE4 were administered to the 5-year-olds with TD and children with history of PPD during the first and second sessions, respectively. The testing protocol for the 8- to 10-year-olds with TD included several literacy measures; therefore, in order to reduce the overall duration of testing, solely IPE4 of the CAPES (Masterson & Bernhardt, 2001) was administered. Furthermore, being 8- to 10-year-olds with typical developmental histories, the children were expected to have mastery of the less phonologically complex words comprising the Profile of the CAPES (Masterson & Bernhardt, 2001). Nevertheless, examiners screened the children's speech production during incidental conversation. As a result, all of the group was considered to demonstrate adultlike speech.

Concerning phonological content coverage (Kirk & Vigeland, 2015) of the CAPES (Masterson & Bernhardt, 2001), the Profile samples 27 monosyllables, 16 disyllables, and three trisyllables, totaling 46 words. The IPE4 samples a wide range of MSWs: 13 disyllables, 29 trisyllables, and 8 four-, 3 five-, and 2 six-syllable words, totaling 55 words. For the single word elicitations, the researchers designed cloze-style prompts, in order to reduce or eliminate the possibility of priming participants for target stress patterns and segments. If the scripted prompt did not elicit the target word, delayed/interrupted imitation (i.e., following a model, the examiner interjected a short indirect request to say the word, before relinquishing the child's turn for production) and, then, immediate imitation were used. The samples were recorded using a Marantz PMD660 digital recorder with a built-in microphone set on automatic level.

Selection of MSWs

Twenty MSWs (see Appendix) from the Profile and IPE4 of the CAPES (Masterson & Bernhardt, 2001) were selected for the current analysis. Selection was on the bases of phonological complexity and word familiarity, the latter to minimize the bias of unfamiliar words that could inflate lexical effects. Word structure complexities included (a) representative lengths, from 1 to 2 ft comprising two to five syllables, and (b) stress patterns, including unstressed word-initial, word-final, and word-medial syllables, in addition to sequences of word-final weak syllables. The included segments were both early and later developing (Smit et al., 1990; Pollock & Berni, 2003), with challenging adjacent and nonadjacent consonant sequences of (a) manner, for example, [±continuant], [nasal], tap, and syllabic; (b) place, for example, coronal–dorsal, dorsal–coronal,

and coronal–labial; and (c) diphthongs and lax vowels. Word familiarity was considered in terms of word frequency and age of acquisition—variables associated with phonological accuracy (Gierut & Morrisette, 2012). In order to avoid parental survey data typically used for age of acquisition, word frequency was chosen to indicate familiarity. Lee (2003) and Morrison, Chappell, and Ellis (1997) have suggested the importance of sourcing words from a large sample of objective child speech data. To achieve these proposed standards and, in addition, include age ranges relevant to the current study, ChildFreq (Baath, 2010) was used to calculate word frequencies.

Transcription and the Mismatch Tally Procedure

To prepare the MSW production data for assignment of mismatch tallies, word productions were narrowly transcribed, following a narrow transcription conventions document (B. M. Bernhardt & Stemberger, 2012). The conventions for assigning mismatch tallies were in accordance with procedures described in Mason et al. (2015), which are reviewed as follows. To begin the MSW mismatch tally procedure, transcript components were aligned with a target row of a spreadsheet, in which columns contained target nonlinear phonological components. Alignment was in absolute order of production, such that columns were added for epenthesis. Mismatch judgments first considered each foot independently for lexical selection influences. No more than one lexical selection mismatch was tallied per foot in order to avoid bias as a result of similarity to words of more than one syllable, for example, *magician* /məˈdʒɪʃn̩/ with *musician* /ˌmjuːˈzɪʃn̩/, or *(hippo)potam(us)* /ˌhɪpəˈpʰɑɾəməs/ with *bottom* /ˈbɑɾəm/, for example, when produced as [ˌhɪpəmˈbɑɾəməs]. When a lexical mismatch was tallied, no further phonological mismatch tallies were assigned unless word structure was affected. The sum of the lexical tallies formed the Lexical subtotal.

Phonological components were next evaluated in turn by level of the hierarchy: stress for the prosodic word[4] and each foot, syllables, consonant and vowel timing units, and segmental features. For each foot, stress mismatches were tallied for shifts of primary stress, or for altered stress patterns as a result of syllable deletion or insertion. For a two-footed word, a maximum of one tally was assigned for prosodic word stress, whether a mismatch occurred in either or both feet. Syllable deletion, insertion, and reduplication were also tallied. Timing unit tallies were assigned if (a) segments were inserted, or deleted without compensatory lengthening, (b) timing units were inserted when a consonant transposed, (c) full vowels were inserted before a syllabic consonant, (d) lax vowels in unstressed syllables appeared as full vowels, and (e) taps were produced as (longer) stops. The latter three were tallied for the reasons that follow. First, for consonants in syllabic versus nonsyllabic contexts, differing acquisition patterns have been suggested (B. H. Bernhardt & Stemberger, 1998). Second, James (2006) has reported developmental lengthening of schwa in MSWs up to age 7 years, implying the existence of a shorter adult timing unit for schwa than for other vowels. Third, the manner of tap production has been distinguished from that of other consonants in terms of its ballistic nature, or speed and trajectory (Ladefoged, 2006; Shockey, 2003). The (Word) Structure subtotal was thus the aggregation of stress, syllable, and timing unit mismatches.

Concerning the final phonological level, segmental features, mismatches were tallied for consonants in terms of place, manner, and laryngeal categories (one tally per category maximum) and, for vowels, with respect to height, backness, and tenseness, in addition to nasalization and roundness as relevant. Nonallophonic and nonassimilatory diacritics, for consonants and vowels, respectively, were also tallied. The combination of feature tallies formed the Feature subtotal. The Whole Word total was therefore the sum of the Lexical, Structure, and Feature subtotals (see Table 1 for example tallies for various productions of *hippopotamus*). For the current study, the aggregations of mismatch tally data were identical to those in the original study (Mason et al., 2015) with the exception that consonant diacritic tallies were included in the Feature subtotal and not solely in the Whole Word total. The original study had demonstrated that variable productions of a word could be differentiated whether or not diacritics were included in the calculations. Combining feature and diacritic tallies, therefore, allowed all segmental mismatches to be represented by the Feature subtotal.

Reliability

Interrater transcription and scoring reliability were completed for independent narrow transcriptions of the Profile and the selected MSWs from the IPE4 of the CAPES (Masterson & Bernhardt, 2001), with data from subsamples balanced for gender. In all cases, point-to-point reliability was between the first author and an SLP graduate of a master's program. First, for the Profile, approximately 15% of the 5-year-olds with TD (Group 1a) and 30% of children with history of PPD (Group 2) were included. Reliability for consonants was 92% and 80%, respectively, considered acceptable levels for narrow transcription (Shriberg & Lof, 1991). Scoring reliability was determined for manually scored data for all of Groups 1a and 2, and was 99% for PCC. Reliability for percent vowels correct (PVC) was not calculated because inconsistency among raters is common.

Second, for the MSWs, 10% of the data was sampled for transcription and scoring reliability. Point-to-point reliability for consonants was 91% for Group 1 (age 5 years) and 83.5% for Group 2 (history of PPD). For Group 3 (age 8 to 10 years), Cohen's kappa was calculated because reliability was between the first author and two other raters, the latter of whom each transcribed half of the original sample. Interrater reliability and consistency were

[4]Defined in terms of nonlinear phonological theory (B. H. Bernhardt & Stemberger, 1998), in which the minimal prosodic word comprises two or more syllables.

Table 1. Tallies for productions of *hippopotamus* /ˌhɪpəˈpʰɑrəməs/ using the multisyllabic word metric.

Production	Lexical effect				Timing units		Features		Subtotals			Word total
	Ft 1	Ft 2	Stress	Syllables	C	V	C	V	Lexical	Word structure	Features	
ˌhɪpəˈpʰɑnəməs[a]	0	0	0	0	1	0	1	0	0	1	1	2
ˌhɪpəˈpʰɑməmɪʂ[b]	0	0	0	0	1	0	3	0	0	1	3	4
ˌhɪpəmˈbɑrəməs[c]	0	1	0	0	1	0	3	0	1	1	3	5
ˌhɪpəˈpʰɑːməs[d]	0	0	2	1	1	0	3	3	0	4	6	10
ˌhɪbəˈnɑnənoʊs[e]	0	1	0	0	1	1	5	4	1	2	9	12

Note. Stress is judged for each foot and for the word overall. /r/ timing unit considered shorter than for C. Lexical subtotal = number of lexical effects. Word structure subtotal = stress + syllable + (C + V) timing unit mismatches. Feature subtotal = C + V feature mismatches. Ft = foot; C = consonant; V = vowel.

[a]/r/>[n], feature tally for [+nasal]. [b]/r/>[m], feature tallies for [Labial] [+nasal]; /s/>[ʂ], feature tally for [-grooved]; lax vowel /ɪ/ accepted variant for /ə/. [c]Third and fourth syllables possibly influenced by word neighbor *bottom* /ˈbɑrəm/; inserted [m], feature tallies for manner, place, and laryngeal. [d]Deleted /r/, feature tallies for manner, place, and laryngeal; deleted /ə/ (timing unit realized as compensatory lengthening), feature tallies for height, backness, and tenseness. [e]Diphthong in word-final syllable possibly influenced by abbreviated word *hippo* /ˈhɪpoʊ/; /p/>[b], feature tally for [+voice]; /pʰ/>[n], feature tallies for [+nasal] [Coronal]; /r/>[n], feature tally for [+nasal]; /m/>[n], feature tally for [Coronal]; inserted [o] (word shape change), feature tallies for height, backness, tenseness, and [+round].

considered fair to good, κ = .668, 95% CI [.607, .669], p = .001 (relative to a criterion of .7; Cohen, 1988).

Next, reliability was examined in independent mismatch coding based on data from 10% of the samples. Point-to-point agreement was 98% or better for mismatch tallies for both groups of typically developing 8- to 10-year-olds (Groups 1a and 3). Reliability and consistency were high for the coding overall (matches and mismatches) and the mismatches alone using the single measure intraclass correlations (ICCs; two-way random effects model): (a) Group 1 (age 5 years), $ICC(2, 1)$ = 0.925, 95% CIs [0.921, 0.930], and 0.819, [0.785, 0.849], p = .001; and (b) Group 2 (history of PPD), $ICC(2, 1)$ = 0.951, 95% CIs [0.948, 0.955], and 0.872, [0.842, 0.896], p = .001.

Statistical Analyses

The purpose of the study was to add to the limited data about phonological accuracy in MSWs for school-aged children with TD and with history of PPD. A whole-word mismatch tally metric was utilized to make three comparisons: (a) the same children with TD, at 5 and 8 to 10 years of age, (b) age-matched 8- to 10-year-olds with TD and history of PPD, and (c) 8- to 10-year-olds with history of PPD and 5-year-olds with TD. The data were entered into SPSS (Versions 19 and 24), and assumptions for normality, linearity, univariate, and multivariate outliers were examined by visually inspecting histograms, Q-Q, box, and scatter plots, in addition to conducting related statistical tests for multicollinearity and sphericity. Unless otherwise stated, assumptions were considered to be met.

Prior to the analysis of the MSW mismatch tallies, PCC and PVC data for the less phonologically complex words (the Profile of the CAPES; Masterson & Bernhardt, 2001) were examined in two nonparametric analyses: (a) the Wilcoxon signed-ranks test (p = .025, two tailed) for dependent samples (Groups 1 and 1a) and (b) the Mann–Whitney U test (p = .025, two tailed) for independent samples (Groups 2 and 3). Medians were tested, and effect size r was calculated.

Data were also examined for average standard score means and group equivalence on two language-related variables: vocabulary comprehension (Peabody Picture Vocabulary Test–Fourth Edition; Dunn & Dunn, 2007) and working memory (Number Repetition Forward or Backward subtests of the CELF-4; Semel et al., 2003). In addition to the Whole Word total tallies, subtotals were also tested, that is, Lexical, Structure, and Features. The data were entered into one-way multivariate analyses (multivariate analysis of variance [MANOVA], $p < .05$), with the probability level adjusted in the situation of unequal variances (p = .01; Field, 2009), and into planned univariate analyses (analysis of variance, $p < .01$), with Bonferroni correction for the number of tests. Mean differences were examined using Wilk's lambda and related F tests, and effect sizes (η_p^2) were calculated. Post hoc discriminant analyses using maximization procedures were also conducted, that is, to determine the linear combination of the MSW subtotals that would maximize the difference in a given comparison, and best account for the variance (Wilk's lambda, with effect size, canonical R^2). Therefore, the sensitivity and specificity of the MSW metric were analyzed in terms of predicting which group participants belonged to. All effect sizes were interpreted according to Cohen's (1988) conventions.

Results

The results of the three comparisons will be presented in the following order: (a) the longitudinal investigation of children with typical phonological development, that is, Group 1a at ages 5 and 8 years; (b) the cross-sectional study of age-matched 8- to 10-year-olds with history of PPD and with TD, that is, Groups 2 and 3, respectively; and (c) the cross-sectional comparison of 5-year-olds with TD and 8- to 10-year-olds with history of PPD, that is, Groups 1 and 2, respectively. The section begins with descriptive

and statistical comparisons on the group-equating, language-related variables and for the preliminary assessment of phonological accuracy, before turning to the main descriptive, statistical, and follow-up analyses.

Group Equating on the Language-Related Variables

To confirm whether the comparison groups had equivalent language-related skills, the relevant groups were first evaluated for equivalence on standard scores for vocabulary comprehension and working memory. Most children scored within the average range as defined in the technical manual for the respective test (see Table 2). For working memory, however, two 5-year-olds and one 8- to 10-year-old with history of PPD scored below average. These participants were retained because of their average performance on the related subtests, Recalling Sentences (Clinical Evaluation of Language Fundamentals Preschool–Second Edition; Wiig et al., 2004) or Segmentation and Elision (Comprehensive Test of Phonological Processing; Wagner et al., 1999), respectively.

For variance–covariance homogeneity, in the longitudinal comparison only, Mauchley's test of sphericity was not met; therefore, Greenhouse–Geisser estimate corrections were applied (Field, 2009). Subsequently, the one-way between-groups MANOVA was without significance in all three analyses, suggesting group equivalence on the language-related measures (see Table 2).

Preliminary Phonological Assessment

Phonological development was evaluated preliminarily in less phonologically complex words using the (Phonemic) Profile of the CAPES (Masterson & Bernhardt, 2001). The Profile samples 27 monosyllables, 16 disyllables, and three trisyllables, totaling 46 words, and was administered to Groups 1 (5-year-olds) and 2 (history of PPD). The Profile was not administered to Group 3 (age matched) in order to reduce the overall duration of the testing protocol, which included several literacy measures. Being 8- to 10-year-olds with typical developmental histories, the children in Group 3 were also expected to have mastery of the words comprising the Profile of the CAPES (Masterson & Bernhardt, 2001).

For the 5-year-olds with typical phonology (Groups 1 and 1a) and the children with history of PPD (Group 2), Table 3 presents phonological accuracy on the Profile of the CAPES (Masterson & Bernhardt, 2001) in terms of PCC (narrow transcriptions) and PVC (broad transcriptions). None of the children demonstrated uncommon clinical substitutions or distortions (e.g., Smit et al., 1990; Pollock & Berni, 2003), with both groups demonstrating generally inconsistent small place shifts for fricatives and affricates and/or vocalization of schwar. Some children in Groups 1 and 2 also distorted the consonantal rhotic, /ɹ/. Judging from the observed mismatch patterns and comparing with available reference data for PCC and PVC (Austin & Shriberg, 1997), all children in Group 1 were considered to have typically developing speech. For Group 2, four children were classified in each of three groups, that is, resolved speech errors (ages 8 to 10 years), questionable residual errors (age 8 years), and residual errors (ages 9 to 10 years), suggesting group heterogeneity as to resolution of PPD.

For the nonparametric tests, results for the 5-year-olds were small and nonsignificant between the whole group (Group 1) and the subgroup that was followed longitudinally

Table 2. Means for the language-related skills.

Measure	Comparison	Group, age[a]	*M*	*SE*	95% CI	*F*	*df*	*p*
Vocabulary (PPVT-4)	Longitudinal	TD at 5	118.27	2.81	[112.61, 123.94]	.189	1, 42	.67
		at 8 to 10	116.55	2.81	[110.88, 122.21]			
	Cross-Sectional 1	PPD 8 to 10	110.75	2.89	[104.77, 116.74]	.736	1, 22	.40
		TD 8 to 10	114.25	2.89	[108.27, 120.24]			
	Cross-Sectional 2	TD 5	117.11	1.70	[113.73, 120.50]	2.28	1, 72	.14
		PPD 8 to 10	110.75	3.86	[103.06, 118.44]			
Working memory (CELF-4[b])	Longitudinal	TD 5	10.18	.50	[9.17, 11.20]	1.63	1, 42	.21
		TD 8 to 10	9.27	.50	[8.26, 10.29]			
	Cross-Sectional 1	PPD 8 to 10	8.75	.64	[7.42, 10.09]	3.36	1, 22	.08
		TD 8 to 10	10.42	.64	[9.08, 11.75]			
	Cross-Sectional 2	TD 5	10.18	.26	[9.67, 10.69]	4.99	1, 42	.03
		PPD 8 to 10	8.75	.59	[7.59, 9.92]			

Note. Longitudinal comparison is for Group1a (n = 22). Cross-Sectional 1 is Group 2 compared with matched Group 3 (ns = 12). Cross-Sectional 2 is Group 1 compared with Group 2 (ns = 62, 12). Means are for standard scores. p = .017. PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition (Dunn & Dunn, 2007), *M* (*SD*) = 100 (15); CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition (Semel, Wiig, & Secord, 2003), subtest *M* (*SD*) = 7 (3); PPD = protracted phonological development; TD = typical development.

[a]Years. [b]Number Repetition Forward and Number Repetition Backward subtests for 5- and 8- to 10-year-olds, respectively.

Table 3. CAPES Profile comparisons for the 5-year-olds with typical development (TD) and 8- to 10-year-olds with protracted phonological development (PPD).

Group	Mean age[a]	*n*	*Mdn* (range) PCC	*Mdn* (range) PVC	Comparisons: Groups 1 and 1a[b] PCC *z* (*p*)	PCC *r*	PVC *z* (*p*)	PVC *r*	Groups 1 and 2[c] PCC *z* (*p*)	PCC *r*	PVC *z* (*p*)	PVC *r*
TD												
1	5.64 (3.08)	62	85.73 (76.99–98.23)	93.76 (83.82–100.00)								
1a	5.60 (3.08)	22	82.68 (74.34–96.46)	95.58 (88.24–100.00)	1.758 (.08)	.27	.257 (.82)	.04				
PPD												
2	9.38 (9.64)	12	89.09 (84.43–93.75)	92.98 (89.76–96.20)					1.248 (.22)	.27	1.968 (.05)	.42

Note. $p = .025$, two tailed; exact significance reported. CAPES = Computerized Articulation and Phonology Evaluation System (Masterson & Bernhardt, 2001); PCC = percent consonants correct for narrow transcriptions; PVC = percent vowels correct for broad transcriptions.
[a]Years (standard deviation in months). [b]Wilcoxon signed-ranks test. [c]Mann–Whitney test; for PCC and PVC, $U = 213.00$ and 175.00, respectively.

(Group 1a). The small difference between Group 1 (5-year-olds) and Group 2 (history of PPD) was also nonsignificant, confirming the classification of some children in the residual errors group; however, the more stringent narrow transcriptions might have overestimated the frequency of error patterns in comparison with (usually) broad transcriptions used for constructing reference data (e.g., Austin & Shriberg, 1997).

Preliminary Analysis of the Phonological Mismatch Tally Variables

To determine whether particular mismatch tally variables were redundant and therefore should be removed from the analyses, for all comparisons, multicollinearity was examined in the variance–covariance matrices. For the correlations of the Structure with Feature subtotals, as well as of the Whole Word total with Structure and Feature subtotals, the strong positive relationships were predictably increasing. These relationships resulted from tallies compounding across the Structure and Feature components for phonological deletions or insertions, in turn contributing to the Whole Word total. Even though the Structure and Feature subtotals were strongly correlated ($r = .68$ to .77, $p = .01$, two tailed), they were retained as separate variables because of their theoretical importance in the nonlinear phonological model. Whereas the Whole Word total was the sum of the subtotals, the large correlation indicated its predictable redundancy ($r = .81$ to .99, $p = .01$, two tailed), such that it was removed from subsequent analyses. Finally, in the comparisons that included 5-year-olds, the significant correlations of the Lexical subtotal with the Structure and Feature subtotals were moderate ($r = .35$ to .46, $p = .01$, two tailed), implying that the Lexical subtotal could be retained in the analyses (Field, 2009). For each comparison, the means for the subtotals, standard errors, and 95% CIs are reported in Table 4.

Five-Year-Olds With TD Followed Between Ages 8 to 10 Years (Group 1a)

To assess the MSW mismatch tally differences, a one-way within-group MANOVA was conducted on the dependent variables: Lexical, Structure, and Feature subtotals. The assumptions were considered met, with exceptions as follows. Mauchley's test of sphericity of the variance–covariances was not met, and consequently, Greenhouse–Geisser estimate corrections were applied in the subsequent analyses (Field, 2009).

Examination of the means revealed that, at 5 years old, the children had consistently more Lexical, Structure, and Feature mismatches than at 8 to 10 years old (see Table 4). The results of the repeated-measures MANOVA for the combined dependent variables indicated a large significant within-group difference, $F(3, 19) = 33.29$, $p = .001$, $\eta_p^2 = .84$. For the follow-up univariate analyses, at the adjusted probability level ($p < .0125$) for the number of tests, all differences were large and significant, $F(1, 42) = 12.58$, $p = .002$; 11.57, $p = .003$; 79.54, $p = .001$; $\eta_p^2 = .38, .36, .79$, for the Lexical, Structure, and Feature subtotals, respectively.

Eight- to 10-Year-Olds With History of PPD (Group 2) Versus TD (Group 3)

To assess the MSW mismatch tally differences, a one-way between-groups MANOVA was conducted on the dependent variables: Lexical, Structure, and Feature subtotals. Levene's test of homogeneity of error variances was significant for the Feature subtotal, $F(1, 22) = 6.65$, $p = .017$. Therefore, the p value for the follow-up univariate tests was reduced to .01 (Field, 2009).

Concerning the mean mismatch subtotals, for both groups, the Lexical subtotal was less than one. Presumably, irrespective of history of PPD, children were familiar with the words. However, Group 2 had more Structure and Feature

Table 4. Means for the multisyllabic word mismatch subtotals for the three comparisons.

Subtotal type	Comparison	Group, age[a]	Subtotal mean	*SE*	95% CI
Lexical	Longitudinal	TD at 5	.56	.13	[0.29, 0.82]
		at 8 to 10	.05	.05	[–0.05, 0.14]
	Cross-Sectional 1	PPD 8 to10	.33	.17	[–0.02, 0.69]
		TD 8 to 10	.42	.17	[0.07, 0.77]
	Cross-Sectional 2	TD 5	.40	.08	[0.24, 0.55]
		PPD 8 to 10	.33	.18	[–0.02, 0.69]
Structure	Longitudinal	TD 5	11.69	1.80	[7.95, 15.43]
		TD 8 to 10	6.10	.70	[4.64, 7.56]
	Cross-Sectional 1	PPD 8 to10	15.50	1.56	[12.26, 18.74]
		TD 8 to 10	6.44	1.56	[3.20, 9.68]
	Cross-Sectional 2	TD 5	11.22	1.05	[9.14, 13.31]
		PPD 8 to 10	15.50	2.38	[10.76, 20.24]
Feature	Longitudinal	TD 5	39.96	3.33	[33.04, 46.88]
		TD 8 to 10	9.81	1.28	[7.14, 12.47]
	Cross-Sectional 1	PPD 8 to 10	46.25	3.45	[39.10, 53.40]
		TD 8 to 10	12.99	3.45	[5.83, 20.14]
	Cross-Sectional 2	TD 5	38.63	2.22	[34.21, 43.06]
		PPD 8 to 10	46.25	5.05	[36.19, 56.31]

Note. Longitudinal comparison is for Group1a (n = 22). Cross-Sectional 1 is Group 2 compared with matched Group 3 (ns = 12). Cross-Sectional 2 is Group 1 compared with Group 2 (ns = 62, 12). TD = typical development; PPD = protracted phonological development.
[a]Years.

mismatch subtotals than Group 3 (see Table 4). In the main analysis, Wilk's lambda indicated a large significant between-groups difference on the combined dependent variables, $\lambda = .30$, $F(3, 20) = 15.71$, $p = .001$, $\eta_p^2 = .70$. At the adjusted probability level ($p < .01$) for the follow-up univariate analyses of variance, the Lexical subtotal difference was nonsignificant, $F(1, 22) = .12$ ($\eta_p^2 = .005$), whereas differences were significant and large for the Structure and Feature subtotals, $F(1, 22) = 16.82$ ($\eta_p^2 = .43$), and 46.52 ($\eta_p^2 = .68$), respectively.

Five-Year-Olds With TD (Group 1) Versus 8- to 10-Year-Olds With History of PPD (Group 2)

A one-way between-groups MANOVA was conducted to test the difference on the MSW dependent variables: Lexical, Structure, and Feature subtotals. For both groups, the mean Lexical subtotals were less than one, suggesting that the words were just as familiar to the younger as to the older children. The mean Structure and Feature subtotals were higher for Group 2 but much less so than in the age-matched comparison (see Table 4). For each mismatch subtotal, however, the overlapping CIs suggested similarity between groups.

The main analysis was conducted using Type III Sums of Squares for the unsystematically unbalanced design, that is, the smaller sample of children in Group 2 resulted from its low incidence relative to the population of 8- to 10-year-olds as opposed to withdrawal for any systematic reasons. Wilk's lambda for the between-groups difference on the combined dependent variables was not significant, $\lambda = .95$, $F(3, 70) = 1.22$, $p = .31$, $\eta_p^2 = .05$, so the data were not tested further.

Follow-Up Discriminant Analyses

To further explore the relationships among the dependent subtotal variables, discriminant analyses were conducted for the two significant comparisons, that is, longitudinal (Group 1a) and cross-sectional (Group 2 vs. Group 3). For the analyses, the eigenvalue revealed one discriminant function that accounted for 100% of the variance; canonical $R^2 = .86, .70$, and significantly differentiated the groups, $\lambda = .26, .30$, $\chi^2(3) = 53.91, 24.82$, $p = .001$. Concerning the loadings of the variables, for the longitudinal comparison, the Feature subtotal loaded highly and twice as much on the function as either the Structure or Lexical subtotals ($r = .78, .27, .35$, respectively). For the cross-sectional comparison, however, only the former was true, that is, the Feature subtotal loaded highly and twice as much on the function as the Structure subtotal ($r = .95, .57$, respectively), but the Lexical subtotal loaded very little ($r = -.048$).

Follow-up classification analyses for the longitudinal and cross-sectional comparisons suggested that the function correctly classified 91.0% and 95.8% of participants, respectively, in their original groupings. Sensitivity of the MSW whole-word metric for classifying children with history of PPD was 100% (12/12). Concerning specificity, accuracy by subgroup with TD was as follows: 86.4% of the 5-year-olds (19/22) and 91.7%–95.5% of the 8- to 10-year-olds (11/12, 21/22), such that three 5-year-olds and one 8- to 10-year-old with TD were classified in the group with history of PPD. For participants of the same age, the results suggested good sensitivity and specificity of the metric, that is, 90% or better (Andersson, 2005). The apparently lower accuracy for classifying 5-year-olds with TD implied the greater variability among the younger children, with some 5-year-olds with

TD expected to have accuracy equivalent to 8- to 10-year-olds with TD.

Discussion

Concerning school-aged children with typical phonological development and with history of PPD, the principle aim of the current study was to document the similarities and differences in the development of phonological accuracy in MSWs, as measured on a whole-word mismatch metric. In contrast, other researchers have counted discrete phonological processes for children with TD up to the age of 7 years (James, 2006) and for preschoolers with PPD (Masso et al., 2016). A secondary aim of the current study was to add evidence to the validity of the whole-word metric, providing a way to quantify a wide range of prosodic and segmental components, and, in turn, to distinguish between children with typical speech development and with history of PPD.

With respect to the principle aim, the main results were as follows: (a) For typically developing children, the frequencies of all three mismatch subtotals, Lexical, Structure, and Features, decreased significantly between ages 5 and 8 years, confirming a developmental progression; (b) for children between 8 and 10 years of age, those with history of PPD demonstrated significantly more Structure and Feature mismatches than an age-matched TD group, but not Lexical mismatches; (c) the frequencies of MSW Lexical, Structure, and Feature mismatches produced by 8- to 10-year-olds with history of PPD were equivalent to those of 5-year-olds with TD.

Concerning the secondary aim, evidence was added to the sensitivity and specificity of the whole-word metric in terms of highly accurate classification of participants in their original groupings (i.e., by age for children with TD and by group for 8- to 10-year-olds with TD and with history of PPD). The expected overlap of some 5-year-olds classified in the 8- to 10-year-old group with history of PPD was also observed.

In order to address the aims, two major hypotheses were proposed regarding between-groups comparisons of MSW accuracy on the metric. The first hypothesis concerned developmental aspects of MSWs in children with TD in a longitudinal comparison. As predicted, the older children (at ages 8 to 10 years) showed significantly fewer mismatches (Lexical, Structure, Features) than the younger ones (at age 5 years). The moderate to large effect sizes for the significant comparisons of the Lexical, Structure, and Feature subtotals suggested that each of these aspects of MSW phonology continued to develop between ages 5 and 8 to 10 years. The finding of moderate differences for the Lexical and Structure mismatches versus large differences for the Feature mismatches suggested the relatively lower feature accuracy of the 5-year-olds. James (2006) used a phonological process frequency analysis to examine word structure and features in MSW production for typically developing Australian English–speaking children between ages 5 and 7 years and reported similar results. That is, deletions and insertions of word structure components (syllables, consonants, and/or vowels) decreased during this period. Patterns related to the production of phonological features (manner, place, and laryngeal/voice) and sequencing of features across segments contained in a word also declined after age 5 years, particularly for words longer than two syllables. Feature mismatches apparently occurred because of challenging sequences in MSWs and traded off with earlier gains in word structure, particularly with stress patterns. After 7 years of age, matches of MSW components on the prosodic and feature levels of the nonlinear phonological hierarchy were synchronized.

The second hypothesis concerned the ability of the metric to identify differences in phonological accuracy in MSWs between 8- to 10-year-olds with TD and those with history of PPD. The larger differences expected in phonological mismatch frequencies were borne out and, in addition, the minimal expected differences in lexical mismatches. It appeared that, with developmental vocabulary growth and familiarity, lexical influences were less important, at least for the words sampled in this study. Inclusion of lexical influences in the mismatch tallies was apparently particularly relevant for the younger children than for the older ones, with TD or with history of PPD. This confirmed that the children with history of PPD were not comprised by lower word familiarity in comparison with their age-matched peers; as expected, however, the group with TD had significantly fewer mismatches in structure and features. In fact, the children with TD showed near mastery of all words on the whole-word metric, such that the predicted moderate differences in structure and feature mismatches were actually large. Whether 8- to 10-year-olds with history of PPD would show higher accuracy than typically developing 5-year-olds was also relevant to the second hypothesis. The predicted moderate differences in lexical mismatches were not confirmed; neither were the predicted small differences in word structure and segmental feature mismatches confirmed. In spite of heterogeneity as to the presence of residual speech sound errors, the 8- to 10-year-olds with history of PPD showed similar mismatch frequencies to those of the 5-year-olds with TD. When quantified on a whole-word measure, the MSW accuracy of children with history of PPD could therefore be expected to be equivalent to that of younger children with typically developing phonology.

The discriminant analyses also contributed information about the sensitivity and specificity of the MSW metric, suggesting good accuracy relative to a minimum classification adequacy criterion of ≥ 86% (Andersson, 2005). The metric that included tallies for Lexical in addition to Word Structure and Feature mismatches accurately classified children according to MSW developmental age group, 5 or 8 to 10 years of age with TD, and with regard to the typicality of phonological development, that is, 8- to 10-year-olds with TD or with history of PPD. The metric was also sensitive to predictable increases in lexical familiarity, in that the most lexical mismatches were made by the typically developing 5-year-olds, whereas there was no difference for 8- to 10-year-olds with TD or with history of PPD. The suggestion was

that phonological difficulty was key for children with history of PPD. The finding of more phonological (i.e., Word Structure and Features) mismatches on average than younger children with TD was also noteworthy, even if not significant. That is, children with other profiles of language disorder have reportedly shown differences in performance in comparison with younger typical children (see Leonard, 2014, for a review).

To date, no other studies of MSWs have quantified phonological structure using a nonlinear framework. James (2006), however, reported PVC and PCC—estimates of linear segmental accuracy. After age 5 years, PVC was 94% or better and, PCC, 90% or better, with 1% growth in each of successive years up to age 7 years. These data suggested little developmental distinction from ages 5 to 7 years and, indeed, phonological mastery of the words. Concerning 4- and 5-year-olds with speech sound disorders, Masso et al. (2016) compared PCC, PVC, and percent phonemes correct for 50 words that were primarily monosyllabic and disyllabic and 30 that were polysyllabic, finding that only PVC distinguished production accuracy. The original study of the MSW metric utilized in the current research (Mason et al., 2015) has also suggested the important contribution of vowels to distinguishing productions with equivalent PCC but very different phonological characteristics, for example, place mismatches for several consonants versus more pervasive stress and foot and syllable structure mismatches because of syllable deletion and reduplication. Tallying vowels provides the opportunity to capture the preservation of syllable nuclei and the interaction of vowels with consonants. The MSW metric appeared to have promise as a method for more narrowly quantifying differences in phonological development of a group of complex words; traditional quantification methodology (PCC, PVC, process analyses) is limited in its sensitivity and specificity relative to MSW development.

Theoretical Implications

For describing phonological acquisition in MSWs, including the multiple error patterns that may occur within a word's production, the current research supports a continuum view over a modular account. The continuum is of acquisition of multiple interacting levels of hierarchically organized (nonlinear) phonological structure (e.g., B. H. Bernhardt & Stemberger, 1998), as opposed to acquisition across stages with overlap of decreasing phonological patterns (e.g., Masso et al., 2016).

To explain the interactions in MSW productions, models of nonlinear phonology can be mapped onto parallel interactive models of language processing. Whereas the phonological hierarchy provides a description of the relationships among the various units of MSWs, models of parallel interactive language processing operationalize the activities within the hierarchy. Mapping is possible because of alignment among a number of constructs of the two types of theories. First, the models account for the link between the lexicon and the prosodic word such that both semantics and phonology could influence MSW output. The relationship of semantic similarity and familiarity to phonological output is thus taken into account. Second, relationships among component domains, or nodes and hierarchical networks, are explained by levels of unit activation and the feed forward–feed backward of information. Third, trade-offs in word structure and segmental accuracy are clarified by the relative activation weights of weaker and stronger prosodic domains. Incorrect weights compromise the amount of cognitive capacity needed to accurately access and select valid phonological components with the appropriate timing. This competition within networks may result in an output that is less complex at one level but more complex at another, as in *giraffe* [dʒəˈɹæːf] produced as [dʒɹæːf], when creation of a [dʒɹ] cluster results from unstressed schwa deletion. The phenomena suggest that more than one factor contributes to phonological mismatches in the output of a word, at least until the relative weights of connections are stabilized in the system.

Clinical and Research Implications

The major strength of the study is the contribution to the quantitative evidence base about MSW phonological production for children of early school age, providing further support for the need to include MSWs in phonological evaluations of school-aged children, broadening the notion of persistent PPD (Lewis et al., 2015). Moreover, even children whose PPD has been considered resolved at school entry may present with ongoing PPD if evaluated in more complex word contexts (i.e., MSWs) than is typically done (James, Van Doorn, & McLeod, 2008; Skahan, Watson, & Lof, 2007). In later elementary school, MSW evaluations of children considered to have residual speech sound errors may demonstrate the pervasiveness of PPD in more phonologically difficult contexts. For one region of Canadian English–speaking children, the study has provided initial criterion reference data for productions of a representative set of culturally relevant MSWs.

An additional strength of the current research is the use of a metric for holistically quantifying MSW production, with utility for research and clinical application. That is, a tool with promising reliability and validity, as well as uniform administration and scoring, is available for capturing the multiple interactions that occur within MSW productions. Another advantage is that quantification of MSW accuracy provides a continuous scale of measurement and, therefore, more power to detect change than arbitrary categorizations (Peterson, Pennington, Shriberg, & Boada, 2009). The sensitivity and specificity of the metric were also considered in relation to phonological development: (a) for children with TD in the early and later elementary school years and, in addition, (b) for children with history of PPD who were heterogeneous as to the resolution of PPD in less phonologically complex words and who demonstrated ongoing protracted phonological production in MSWs.

There are a number of implications and needs for future research. First, future research will need to expand the sample in terms of size, geographic location, age groups,

and groups by developmental language level (monolingual and multilingual). Ideally, to verify that participants' language development is within the average range, skills could be directly and comprehensively assessed at the time of the study. Instead, for the current study, clinician's reports of previous documentation were relied on to reduce the overall duration of testing. Furthermore, for some children, although language development appeared typical prior to the study, difficulty in later language acquisition might become apparent if assessed.

Second, the current study included words with a variety of morphemic compositions, that is, monomorphemic or multimorphemic. Words were sometimes compound nouns (i.e., *watermelon*, *cash register*), verb derivations (e.g., *invitation*), or inflected verbs or pluralized nouns (e.g., *explodes*, *balloons*). Inclusion of these words provided opportunities to observe the interactions of more phonological components within the nonlinear hierarchy. It was presumed that children were not yet aware of derivational relationship between some of the words. However, in future studies, a more balanced list of words relative to derivational morphology is warranted. Another consideration concerns priming effects. Priming was possible as a result of the immediate or preceding word elicitation prompts or from a preceding word in the elicitation order. In future research, the influence of priming effects could therefore be considered in relation to variability in MSW production, but in general, attention to reducing such effects should contribute to study designs.

Finally, for the current sample, in addition to the quantitative metric, the various phonological mismatch patterns could be further analyzed in a descriptive way, providing more qualitative information about the children's overall outputs for the set of MSWs. Up to the present, the evidence base concerning phonological pattern analyses in MSW phonology has been limited primarily to studies of Australian English–speaking children (James, 2006; Masso et al., 2016) with typically developing phonology (ages 3;0 to 7;11) and preschoolers with history of PPD (ages 4;0 to 5;5). The current research, however, was the first to examine phonological accuracy in MSWs in Canadian English–speaking school-aged children with TD and with history of PPD (ages 5 and 8 to 10 years).

Importantly, further development of the psychometric properties of the MSW metric is warranted. There are additional aspects of reliability and validity yet to be established, for example, test–retest reliability, internal consistency, and predictive validity. Second, concerning lexical effects, it would potentially be helpful to establish specific guidelines to distinguish the effects of word neighborhood activations from those of the phonological system. The ICC indicated high coding reliability overall; however, reliability of lexical coding could not be specifically calculated in the current study because the proportion of tallies was small. Larger studies are needed to consider the confounding effects of lexical and phonological tallies because the respective subtotals require independence in order to be additive. If lexical considerations were to be excluded from the metric, however, there would be a loss of valuable information. Additional development of the metric may further elucidate the interaction of lexical and phonological factors.

In conclusion, for school-aged children with history of PPD, MSWs are a necessary part of evaluation for research and clinical practice, to determine if PPD is actually ongoing and, if so, to what extent. For the phonological mismatch metric used in the current study, the Whole Word total has the ability to detect weaknesses and track progress in global MSW phonological production, whereas the subtotals can be used to examine subcomponents of MSW processing, that is, Lexical influences, Word Structure, and segmental Features. In spite of the many word-specific interactions in MSWs, the subtotals might inform phonological patterns for a subgroup of words with certain components in common. The information gained from the various sums within the MSW metric is applicable to intervention planning for children with history of PPD and, in addition, to future studies of MSW phonological development. This new knowledge may help prevent the exclusion of deserving children from appropriate contextually relevant phonological intervention in order to enhance their success at school and their options in life more broadly.

Acknowledgments

The research was part of the author's doctoral dissertation. Funding for this project was provided by The University of British Columbia Hampton Fund and Social Sciences and Humanities Research Council of Canada (410-2009-0348). The author wishes to thank the participants and their families and the doctoral advisors and graduate student research assistants who contributed their time and expertise. Local colleagues are also thanked for their critiques of the article.

References

Andersson, L. (2005). Determining the adequacy of tests of children's language. *Communication Disorders Quarterly, 26,* 207–225.

Austin, D., & Shriberg, L. D. (1997). *Lifespan reference data for ten measures of articulation competence using the speech disorders classification system (SDCS)*. Madison: University of Wisconsin—Madison, Waisman Center, Language Analysis Laboratory.

Baath, R. (2010). ChildFreq: An online tool to explore word frequencies in child language. *LUCS, Minor, 16*. Retrieved from http://childfreq.sumsar.net/

Bernhardt, B. H., & Stemberger, J. P. (1998). *Handbook of phonological development: From the perspective of constraint-based nonlinear phonology*. San Diego, CA: Academic Press.

Bernhardt, B. M., & Stemberger, J. P. (2012). Translation to practice: Transcription of the speech of multilingual children. In N. Muller & M. Ball (Eds.), *Multilingual aspects of speech sound disorders in children* (pp. 182–190). North York, ON, Canada: Multilingual Matters.

Bernhardt, B. M., Stemberger, J. P., & Charest, M. (2010). Intervention for speech production in children and adolescents: Models of speech production and therapy approaches. Introduction to the issue. *Canadian Journal of Speech-Language Pathology and Audiology, 34,* 157–167.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
Dawson, J. I., & Tattersall, P. J. (2001). *Structured Photographic Articulation Test II Featuring Dudsberry* (SPAT-D II). DeKalb, IL: Janelle.
Dubasik, V. L., & Ingram, D. (2013). Comparing phonology of dyads of children with typical development and protracted development. *Clinical Linguistics & Phonetics, 27,* 705–719.
Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test–Fourth Edition* (PPVT-4). Toronto, ON, Canada: Psycan.
Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage Publishing.
Gierut, J. A., & Morrisette, M. L. (2012). Age of word acquisition effects in treatment of children with phonological delays. *Applied Psycholinguistics, 33,* 121–144.
Hodson, B. W. (2004). *Hodson Assessment of Phonological Patterns–Third Edition*. Austin, TX: Pro-Ed.
Ingram, D. (2002). The measurement of whole-word productions. *Journal of Child Language, 29,* 713–733.
James, D. G. H. (2006). *Hippopotamus is so hard to say: Children's acquisition of polysyllabic words* (Unpublished doctoral dissertation). The University of Sydney, Sydney, Australia.
James, D. G. H., Ferguson, W. A., & Butcher, A. (2016). Assessing children's speech using picture-naming: The influence of differing phonological variables on some speech outcomes. *International Journal of Speech-Language Pathology, 18,* 364–377.
James, D. G. H., Van Doorn, J., & McLeod, S. (2008). The contribution of polysyllabic words in clinical decision making about children's speech. *Clinical Linguistics & Phonetics, 22,* 345–353.
Kehoe, M. M. (2001). Prosodic patterns in children's multisyllabic word productions. *Language, Speech, and Hearing Services in Schools, 32,* 284–294.
Kirk, C., & Vigeland, L. (2015). Content coverage of single-word tests used to assess common phonological error patterns. *Language, Speech, and Hearing Services in Schools, 46,* 14–29.
Ladefoged, P. (2006). *A course in phonetics* (5th ed.). Boston, MA: Wadsworth.
Lee, C. J. L. (2003). Evidence-based selection of word frequency lists. *Journal of Speech-Language Pathology and Audiology, 27,* 170–173.
Leonard, L. B. (2014). *Children with specific language impairment* (2nd ed.). Cambridge, MA: MIT Press.
Lewis, B. A., Freebairn, L., Tag, J., Ciesla, A. A., Iyengar, S. K., Stein, C. M., & Taylor, H. G. (2015). Adolescent outcomes of children with early speech sound disorders with and without language impairment. *American Journal of Speech-Language Pathology, 24,* 150–163.
Mason, G. K., & Bernhardt, B. M. (2014). The impact of protracted phonological disorders on literacy outcomes in children: A meta-analysis. In P. A. Ysunza (Ed.), *Speech, language and voice pathology: Methods, challenges and outcomes* (pp. 49–116). New York, NY: Nova.
Mason, G. K., Bérubé, D., Bernhardt, B. M., & Stemberger, J. P. (2015). Evaluation of multisyllabic word production in Canadian English-or French-speaking children within a non-linear phonological framework. *Clinical Linguistics & Phonetics, 29,* 666–685.
Masso, S., McLeod, S., Baker, E., & McCormack, J. (2016). Polysyllable productions in preschool children with speech sound disorders: Error categories and the Framework of Polysyllable Maturity. *International Journal of Speech-language Pathology, 18,* 272–287.
Masterson, J., & Bernhardt, B. (2001). *Computerized Articulation and Phonology Evaluation System (CAPES)*. Toronto, ON, Canada: The Psychological Corporation.
Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 50*(A), 528–559.
Newcomer, P. L., & Hammill, D. D. (1997). *Test of Language Development–Primary: Third Edition* (TOLD-P:3). Austin, TX: Pro-Ed.
Peterson, R. L., Pennington, B. F., Shriberg, L., & Boada, R. (2009). What influences literacy outcome in children with speech sound disorder? *Journal of Speech, Language, and Hearing Research, 52,* 1175–1188.
Pollock, K. E., & Berni, M. C. (2003). Incidence of non-rhotic vowel errors in children: Data from the Memphis Vowel Project. *Clinical Linguistics & Phonetics, 17,* 393–401.
Presson, N., & MacWhinney, B. (2011). The competition model and language disorders. In J. Guendouzi, F. Loncke, & M. J. Williams (Eds.), *Handbook of psycholinguistic and cognitive processes* (pp. 31–48). New York, NY: Psychology Press.
Rvachew, S., & Brosseau-Lapré, F. (2012). *Developmental phonological disorders: Foundations of clinical practice*. San Diego, CA: Plural.
Semel, E., Wiig, E. H., & Secord, W. (2003). *Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4)*. San Antonio, TX: Pearson.
Shockey, L. (2003). *Sound patterns of spoken English*. Oxford, United Kingdom: Blackwell.
Shriberg, L. D., & Lof, G. L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics & Phonetics, 5,* 225–279.
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,* 246–259.
Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The Iowa Articulation Norms Project and its Nebraska replication. *Journal of Speech and Hearing Disorders, 55,* 779–798.
Storkel, H. L., & Morrisette, M. L. (2002). The lexicon and phonology: Interactions in language acquisition. *Language, Speech, and Hearing Services in Schools, 33,* 24–37.
Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processing (CTOPP)*. Austin, TX: Pro-Ed.
Wheeler, D., & Touretzky, D. S. (1997). A parallel licensing model of normal slips and phonemic paraphasias. *Brain and Language, 59,* 147–201.
Wiig, E. H., Secord, W., & Semel, E. (2004). *Clinical Evaluation of Language Fundamentals Preschool–Second Edition* (CELF Preschool-2). San Antonio, TX: Pearson.
Williams, K. T. (2007). *Expressive Vocabulary Test–Second Edition* (EVT-2). San Antonio, TX: Pearson.

Appendix

Multisyllabic Words Analyzed With the Whole-Word Metric

Word	IPA target	Word	IPA target
computer	kʰəmˈpʰjurɚ	magician	məˈdʒɪʃn̩
gorilla	gəˈɹɪlə	animal	ˈæ̃ːnəml̩
explodes	ˌɪkˈsploʊdz	skeleton	ˈskɛlətn̩
guitar	gəˈtʰaːɹ	mosquito	məˈskiːroʊ
giraffe	dʒəˈɹæːf	invitation	ĩnvəˈtʰeɪʃn̩
balloons	bəˈlũːnz	alligator	ˈæːləˌgeɪɾɚ
hospital	ˈhɑːspɪɹl̩	watermelon	ˈwɑːɾɚˌmɛln̩
vegetable	ˈvɛdʒtəbl̩	cash register[a]	ˈkʰæʃˌɹɛdʒɪstɚ
electric	əˈlɛktɹɪk	thermometer	θɚˈmɑːmɚɾɚ
umbrella	ˌʌ̃mˈbɹɛlə	hippopotamus	ˌhɪpəˈpʰɑːɾəməs

Note. IPA = International Phonetic Alphabet.

[a]Although spelled as two words, younger children may have a holistic single and/or compound word representation, as opposed to the derivational relationship.