Identification of protracted phonological development across languages - The Whole Word Match and basic mismatch measures

Barbara May Bernhardt[1], Joseph Paul Stemberger[1], Daniel Bérubé[2], Valter Ciocca[1], Maria João Freitas[3], Diana Ignatova[4], Damjana Kogošek[5], Inger Lundeborg Hammarström[6], Thora Másdóttir[7], Martina Ozbič[8], Denisse Perez[9], and A. Margarida Ramalho[3]

[1] University of British Columbia, Canada
[2] University of Ottawa, Canada
[3] University of Lisbon, Portugal
[4] University of Sofia, Bulgaria
[5] University of Ljubljana, Slovenia
[6] University of Linköping, Sweden
[7] University of Iceland, Iceland
[8] LOGOS-private SLT practice, Sežana, Slovenia; Institute for Maternal and Child Health (IRCCS) La Nostra Famiglia, Udine, Italy
[9] University of Valparaíso, Chile

Please address all correspondence to Barbara May Bernhardt, bernharb@mail.ubc.ca, 2177 Wesbrook Mall, Vancouver, BC, V6T 1Z3, Canada, 1-778-847-4381

Abstract

Identification of children with protracted phonological development (PPD) involves both comparison of a child's phonological skills with those of peers (typically, age-matched), and consideration of other factors concerning the child and his or her communicative context (e.g., psychosocial, oral-motor, auditory perceptual, cognitive, environmental). Relative to comparison, norm-referenced tests are often used in clinical contexts, with the comparisons only as valid and reliable as the sample size and type, and content coverage of the language's or dialect's phonology (Kirk & Vigeland, 2014). Global measures of 'accuracy', e.g. Percent Consonants Correct (Shriberg et al., 1997) or Phonological Mean Length of Utterance (Ingram, 2002) have also been used, and are becoming more streamlined, with phonological analysis programs such as Phon (www.phon.ca). The current paper explores the potential utility of two measures that may be applied in identification of protracted phonological development (PPD) in children: (1) a simple accuracy measure, Whole Word Match (WWM: yes-no congruence of adult and child productions of a word); and (2) for more borderline cases (not clearly typically developing, TD, or with PPD), a composite mismatch measure (based on consonant deletion, vowel changes, consonant substitutions). Data are presented for eight languages: Germanic (German, Icelandic, Swedish); Romance (Canadian French, European Portuguese, Granada Spanish); and South Slavic (Bulgarian, Slovenian). The data comprise phonetically transcribed single word elicitations of about 100 words per sample by child (full lists) and for all but German, subsets of the full lists (screening probes). The word lists and transcription conventions were generated jointly by native speakers and team leaders in order to enhance reliability and comparability across languages. The computer program Phon (Rose & MacWhinney, 2014; Hedlund & Rose, 2018), spreadsheets and statistical programs supported analysis, which included a Bayesian analysis for Bulgarian as a preliminary statistical exploration of WWM. Screening and full word lists were compared within language for groups and individual children. Results showed overall relevance of Whole Word Match as an identifier of PPD across languages (agreement with the original TD/PPD classification for 325/333 children and similar levels of WWM by age across languages. Mismatch measures disambiguated most of the few borderline cases. The chapter concludes with implications for future research and clinical applications.

Keywords: speech sound disorders, articulation disorders, phonological impairment, phonological disorders, typical development, phonological errors, phonological screening

Comment [QZ1]: Abstract for Elena Babatsouli, but will not appear in book

Introduction

Accurate identification of speakers with protracted phonological development (PPD) can be challenging, because of the variability in development and differences in community expectations across linguistic and cultural environments. However, to the degree that it is possible, clinicians and researchers need valid, reliable and efficient methods for such identification. One or more quantitative measures are typically used, with a speaker's skills being compared with those of peers (usually age-matched). However, other factors are also considered concerning the speaker and his or her communicative context (e.g., psychosocial, oral-motor, auditory perceptual, cognitive, environmental). In this chapter, we focus on quantitative data for children speaking one of eight languages (Germanic, Romance and South Slavic) but underline at the outset that other factors are also key considerations. The following sections review existing quantitative measures briefly as background for the measures investigated here: an accuracy measure (Whole Word Match), and a basic mismatch (error) measure.

Global Measures in Identification of PPD

One major quantitative consideration in any developmental classification is how a person performs relative to some criterial level for his or her (mental or chronological) age. For phonology, speech-language therapists (SLTs), and often researchers, generally employ norm-referenced commercially published articulation/phonology tests to identify speakers with PPD. However, such tests tend to be limited due to: (1) a focus on consonants without respect to word length, word structure or phonotactics; and (2) insufficient norm-referencing in terms of participant demographics (Kirk & Vigeland, 2014).

Some researchers have recommended incorporating global measures into the identification process, either focusing on accuracy or mismatches. For example, Shriberg and colleagues (e.g. Shriberg, 1997; http://www2.waisman.wisc.edu/phonology/pubs-tech.html) have provided criterion reference data for a set of accuracy measures, i.e. Percent Consonants Correct, Percent Phonemes Correct, Percent Vowels Correct, etc. Some of their measures ("PCC-Adjusted") discount typical developmental patterns such as lisping in younger children but none of them incorporate all aspects of the phonological system into one measure. Ingram and colleagues designed accuracy measures that integrate vowels and certain aspects of word length into the analysis, e.g., Phonological Mean Length of Utterance (PMLU, e.g., Ingram, 2002; Arias & Lleó, 2015), utilizing arbitrary weighting in order to account for differences in complexity of different words. Focusing on mismatches, Preston and colleagues (2011) also utilized arbitrary weighting in an attempt to account for the different impact of various mismatch types on intelligibility. Recently, Van Borsel and D'haeseleer (2018) established preliminary normative data for a Phonological Process Density Index for Dutch-learning children, i.e., the relative proportion of phonological processes in a sample (described originally in Edwards, 1992). Mason et al. (2015) and Mason (in press) utilized a measure that tallies mismatches non-arbitrarily across all aspects of the phonological system for multisyllabic words: word structure, consonants, vowels, phonological features, sequences. All such measures provide information not available in standard phonology tests and have their individual strengths and utility. However, they all have limitations. Some focus only on one aspect of the phonological system (e.g., PCC, PVC; multisyllabic words), while others utilize arbitrary weighting (Preston et al., Ingram and colleagues) or are subject to differences in definition (phonological processes). All

have limited norming. Furthermore, by hand, most of the measures are fairly labour-intensive. Phon, a computer program, has begun automating more of the measures (Rose & MacWhinney, 2014: Hedlund & Rose, 2018), but Phon presently remains a research tool. Thus, the previous measures are less likely to be used in clinical contexts.

Rationale for the Whole Word Match

In an ongoing crosslinguistic study of over 15 languages, Bernhardt, Stemberger and colleagues (Bernhardt & Stemberger, 2017) have been documenting the phonological skills of children with PPD aged from 3 to 5 years and, where funds allow, those of control groups. In each country, researchers individually classify the children as PPD or typically developing (TD) based on their test performance and other factors (including parent or teacher report). In order to compare data across languages, the question arose as to whether the same criteria are being applied for participant classification. In order to collect similar data across languages, single-word phonological naming tasks have been constructed of about the same length (c. 100 words per child) and content coverage (considering word structure and phonotactics plus consonants and vowels of each language). In addition, similar phonetic transcription conventions and data analysis procedures have been utilized across the languages (Bernhardt & Stemberger, 2012). Participants with PPD have been recruited who have no other major developmental considerations (reduced sentence length being acceptable). However, the classification criteria for PPD could still differ across countries. The question was whether there might be a comprehensive quantitative measure of the phonological system that could be used to compare participant samples. Looking at previous research, Schmitt, Howard and Schmitt (1983) described use of a whole-word accuracy metric in evaluation of speech samples for English, a measure which was developed further by Ingram (2002). Because a whole-word accuracy measure is relatively simple to automate in spreadsheets, the Percent Whole Word Match (WWM) was introduced into the project as one way to compare samples within languages (between TD/PPD) and across languages. (Phon did not incorporate measures such as WWM, PCC, or PMLU until the crosslinguistic study was well underway.) The advantages were: (a) that the metric included vowels, word structure and stress in addition to consonants, i.e. the whole word, enhancing content validity; (b) avoidance of arbitrary assignment of weighting measures; and (c) quick and efficient calculation. In essence, the child's production is compared with the adult production in terms of an exact match.

The questions for the research study (addressed in the current chapter) were: (1) whether children's scores on Whole Word Match agreed with the researchers' classifications of TD versus PPD; and (2) if there was ambiguity (borderline cases), whether a mismatch metric might resolve this ambiguity. As a side benefit, because the Whole Word Match measure is time-efficient, if it were sufficiently reliable, it might be then clinically applicable, especially in contexts where the clinician is asked to evaluate the child's phonology of a language s/he does not speak.

Method

Languages and participant samples

Results are presented in this chapter for groups of children speaking one of eight languages: three Germanic (German, Icelandic, Swedish), three Romance (Canadian French-Manitoba; European Portuguese, Granada Spanish); and two South Slavic (Bulgarian, Central

Slovenian). Participant numbers and age groups vary somewhat, but all languages show data for 4-year-olds, with 3-year-old data unavailable for Slovenian and Swedish, and 5-year-old data presented only for Spanish and Bulgarian. The Portuguese sample has only one participant designated with PPD in each age group, and 27 to 30 with typical development. (See Table 1.) For all languages except Slovenian, TD/PPD classification was based on parent, teacher or SLT referral, supported by an existing test for the language in some cases (Spain, Portugal, Bulgaria) and finally confirmed by the researcher's judgment (all SLTs) in response to the child's performance on the assessment tool for the crosslinguistic project. In Slovenia, all children in the preschools were tested, and children showing notably low scores on a variety of phonological measures were designated as PPD for the study.

Insert Table 1 about here

Speech Data

For all languages except German (full word list only), data are presented for both a screener word list and a full word list by speaker. Children all received the full list of words for evaluation. Later, a screening list was extracted from the full list for each language that matched the full list in proportion of various word structures and content coverage of consonants and vowels. The two scores had two sub-purposes: (1) to see whether the screening lists would give similar results to the full list, i.e. were reliable as screeners; and (2) for possible clinical application, to alert SLTs as to which children might require elicitation of the full list for either diagnostic or intervention purposes. For Icelandic, different but overlapping tests were used for the TD and PPD samples. However, only words in common on the two Icelandic lists were compared. (Statistical comparisons of the children's performance on the comparison list and full lists showed no significant differences in match levels within each age and developmental group. justifying the use of the comparison lists.) (See Table 1 for numbers of words per speech sample.)

Whole Word Match was calculated first using Phon and then double-checked by one or more humans, using the same rules of coding until 100% consensus was reached. Small details of narrow phonetic transcription such as partial devoicing, word-final aspiration of stops, or slight fronting or backing of consonants were ignored throughout, and in an additional analysis, degrooving of sibilants ('frontal lisping') was also ignored, with the perspective that preschoolers' sibilants are not adult-like in any case due to anatomical restrictions. Transcription reliability had been done previously for all the samples and was either 100% by consensus or over 90%, with small details of narrow phonetic transcription sometimes differing across countries.

The first author determined the cut-off criteria for PPD for each sample, conferring with the researcher who collected the data. The means and standard deviations for the sample were taken into account, but because individual scores can skew measures in small samples, graphs were also inspected visually, with the goal of including as many children as possible within each of the two categories. Where children's WWM scores were near the cut-off line, these are indicated by hash striped bars (see Appendix: Figures). Following the analysis of the full test data, if a child was even minimally above the borderline and was originally designated as TD, s/he was considered TD. It is important to emphasize that the data are preliminary descriptive measures

for these languages and are based on small numbers of children. The investigation is presented as a proof of concept more than a set of firm criterial levels for WWM.

As a harbinger of potential future investigations with WWM, a Bayesian statistical analysis was applied to the 4-year-old Bulgarian data. The methodology for this analysis is explained along with the results for that analysis in a section on future directions in the Discussion.

Where ambiguity in identification followed administration of the full list of words, a composite mismatch analysis was conducted in an attempt to disambiguate borderline cases. A basic yes-no tally for whole word match fails to take into account all the possible mismatches in a word, i.e. a word with one mismatch is scored "0" as is a word with two or more mismatches. A finer-grained analysis has potential to disambiguate borderline cases with only yes-no WWM tallies. The mismatch analysis tallied basic changes to the word: consonant deletions, consonant substitutions, epenthesis (of consonants or vowels) and vowel changes, the latter including vowel deletions (i.e. syllable deletions), diphthong reduction and vowel substitutions. The perspective follows tenets of nonlinear phonological theory that word structure is as important as segments: thus, deletions and epenthesis, which affect structure, were considered separately from substitutions for consonants (Bernhardt & Stemberger, 1998). Vowels showed fewer changes overall and thus all mismatches were tallied within a "vowel changes" category, whether they affected word structure or involved feature changes. The composite mismatch measure, although hand tallied, was partially automated through outputs of Phon in spreadsheets.

Results

The major sections of the results describe WWM by age groups, allowing comparability of WWM by age. (German 3- and 4-year-old data are presented in one figure because there were no screener data.). The Appendix presents WWM data in figures 1-15 for all the languages in age order, and within age groups, by language family (and in alphabetical order). For German, Bulgarian, French, Slovenian and Icelandic, WWM is displayed with and without degrooving ('lisping') considered a match, within a stacked column. For Swedish and Portuguese, small deviations in grooving were ignored for this analysis. For Spanish, adults in Granada vary within and across speakers in use of grooved versus ungrooved coronal fricatives and thus, both degrooved and grooved sibilants were considered matches and no contrast is provided. Table 2 summarizes the cut-off criteria for PPD suggested by the data by age. The final section of the results provides a Spanish example of a composite mismatch analysis for disambiguation of classification (TD/PPD), and reports briefly on the mismatch analyses for other languages.

Insert Table 2 about here

Overall: Whole Word Match

Inspection of the figures and the means and standard deviations for the screeners and full word lists revealed the following:

1. Overall, the children classified as PPD had lower WWM levels than the children designated as TD (e.g., a Mann Whitney *U* for Spanish of WWM, $p < .001$). However, our interest in this study

was not in group comparisons; the goal was to ascertain whether each individual's WWM score accorded with his or her designation as TD/PPD, irrespective of group differences.
2. WWM varied across the age range, i.e., there was not a clear increase in WWM by age in months.
3. The WWM scores for the screening and whole word lists were very similar in means/standard deviations and for individuals, suggesting split-half reliability for the full test.

Three-Year-Olds: WWM cut-off criteria for PPD

Figures 1 to 6 in the Appendix show data for German, Icelandic, Canadian French, Granada Spanish, European Portuguese and Bulgarian 3-year-olds respectively. WWM cut-offs for PPD varied between about 15%-45% across the languages.

For Granada Spanish, Bulgarian and Icelandic, a 40% WWM cut-off criterion differentiated most of the children as TD or PPD. For Icelandic (Figure 2), one TD child's scores on the screener and full list were notably lower than those of the other TD children, but were still much higher than scores of children designated as PPD; furthermore, if degrooved sibilants were counted as matches, his WWM scores were in the TD range. For the Spanish screener list (Figure 5), there were several children's scores at around 40% WWM, with the PPD all below this level, however, and three children designated as TD very close to this cut-off. WWM for the full list showed only one TD child still slightly below this level (false negative?) We return to his score in the mismatch analysis at the end of the Results.). For Bulgarian (Figure 6), results were similar to the Spanish. Children designated as PPD were below the 40% WWM level, but three children designated as TD were slightly below 40%, unless, as with Icelandic, degrooving was discounted as a mismatch in which case their scores were above 40%. (As noted above, Granada Spanish allows degrooved sibilants in the adult language and thus, degrooved sibilants were automatically built into the Spanish analysis as a match.)

Canadian French (Figure 4) showed slightly higher cut-off scores for the screener and full lists, 45% and 50% respectively. There were two borderline cases on the screener, one in the PPD and one in the TD groups, both showing small increases in WWM with degrooving accepted. On the full list, the TD child was clearly in the TD range, but the PPD child remained borderline (false positive?), especially given acceptance of degrooving as a match. The participant sample was smaller than the other groups (13 total).

European Portuguese and German both showed lower cut-off scores than the other languages, 15% and 25% for European Portuguese, 25-30% for German (full list only). For German, the three borderline TD cases would have exceeded the 30% criterion if degrooving was considered a match, but one child designated as PPD (by more than one clinician) was also above the 30% cut-off (false positive?). For European Portuguese, the one child with PPD had no matches, but 4/27 of the children designated as TD had scores under 15% on the screening list, and 1/27 below 25% on the full list (false negative?), with a higher score than on the screening list, i.e. TD8 at 42 months. No degrooving analysis was conducted for the Portuguese, and so any contributions of the degrooving analysis are unknown at this time.

Overall, WWM cut-off levels were fully consistent with the TD designation on the screening list for 55/67 children with the strictest criteria, and for 63/67 children if degrooving was considered close enough (irrelevant, however, for the three children speaking Granada Spanish, and unknown for Portuguese). For the screener, WWM cut-offs identified 39/40 of the children originally classified as PPD as PPD (with one exception for Canadian French). For the

full list, results were similar, with an increase in concordance for two of the three TD Spanish children who were no longer classified as borderline PPD. Additionally, for the German children, WWM cut-off identified 1/9 PPD children as borderline TD, and 2/9 TD children as borderline PPD, although for the latter, only if degrooving was considered a mismatch. Thus, in general, WWM did not completely conform to the researchers' original classifications, but with degrooving considered a match, the congruence was relatively high (102/107 hits).

Four-Year-Olds: WWM cut-off criteria for PPD

Figure 1 and Figures 7 to 13 in the Appendix show 4-year-old data for the eight languages: German, Icelandic, Swedish, Canadian French, Granada Spanish, European Portuguese, Bulgarian and Slovenian.

The cut-off criteria ranged from 40-65% WWM for the screening and full lists for 4-year-olds across languages. On the screening list, Icelandic, Swedish, Bulgarian and Canadian French showed a 50% cut-off. With this cut-off level, only one child originally designated as PPD appeared to be TD (Bulgarian false positive), and four children originally designated as TD appeared to be borderline PPD (false negatives). However, if degrooving was ignored, all four of those children were within the TD range for WWM. Both Spanish and Slovenian showed a clear split between TD and PPD (65%, Spanish; 60% Slovenian). For European Portuguese at 40% WWM, 4/30 TD children would be considered borderline PPD (false negatives) on the screener. The full list showed similar patterns, with only two of the Portuguese TD children having borderline scores, however. The German list showed similar levels of agreement, with 2/10 TD children identified as borderline PPD (false negatives) unless degrooving was ignored. WWM did identify all of the children designated as PPD, however, for both German and Portuguese (no false positives).

Overall, then, WWM cut-off criteria for 4-year-olds agreed with the original designation, more so than for the 3-year-olds. Only one child with an original PPD designation (Bulgarian) was identified as TD, and the TD borderline cases were classifiable as TD if degrooving was ignored.

5-year-olds: WWM cut-off criteria for PPD

There were sufficient data for an analysis of Spanish and Bulgarian 5-year-old data (Appendix: Figures 14-15). The cut-off levels for PPD in Spanish were 80% WWM for the screener and 75% for the full list, with two children being ambiguous as to designation on the screener (one PPD, one TD). The full list supported the original designations (PPD as PPD, TD as TD). For Bulgarian, results were similar but there was an overall lower level of accuracy (55%, similar to the 4-year-old data). Most of the Bulgarian 5-year-olds showed degrooving, however, and thus the relevance of degrooving for classification of TD/PPD was unclear. Without degrooving, all children were accurately classified by WWM as per their original designation with the exception of one Bulgarian TD child (who remained borderline).

Disambiguation of Borderline Cases

As noted previously, mean and individual scores for the screener and full lists were close overall (generally a mean difference of 5% WWM or less, with Portuguese and Bulgarian showing a 10-12% difference for only some comparisons). Where WWM was ambiguous on the screener, the full list showed greater concordance with the original classification as TD/PPD in 13/25 cases where it was relevant (it was not relevant for Spanish, Swedish and Slovene 4-year-olds where there was no ambiguity). If degrooving was further allowed for a match, there remained eight children for whom WWM scores on the full list did not accord with the original classification. For those eight cases, a composite mismatch analysis was undertaken. One language example is given here, with a brief account of findings for the other languages.

Spanish example: Disambiguation through composite mismatch analysis

Table 3 provides an example of the composite mismatch analysis for Spanish 3-year-olds, showing the proportions of mismatches relative to total number of words overall, consonant deletion, consonant substitution, vowel changes, and epenthesis. Data are presented for three children scoring below but close to the 40% cut-off for PPD in Spanish 3-year-olds. As can be seen, the child originally designated as TD, had less than one mismatch per word, and a low deletion score, whereas the two originally designated as PPD had more than one mismatch per word and a higher deletion or substitution score. Thus, the children are distinguishable in mismatch proportions and types. For all three cases, the clinical decision might be to review their progress later on, due to their age and proximity to typical timelines; a shorter review period might be anticipated for the two showing higher probability of PPD.

Insert Table 3 about here

Similar results were found for the other languages using a mismatch analysis. For one German 3-year-old designated as PPD but scoring in the WWM TD range, the mismatch score was less than one mismatch per word (.78), i.e., in the range of TD according to the Spanish mismatch analysis above and a possible further indication of a false positive or some other reason for the PPD designation. For the Bulgarian 4-year-old originally classified as PPD but scoring in the WWM TD range, the mismatch analysis had a similar outcome; at age 4, the child had a mismatch ratio of .51, i.e. an average of one mismatch in every two words, a proportion that did appear to be consistent with typicality in the other languages (see also the Bayesian analysis in the Discussion which further addresses that child's performance, in the Discussion). Thus, a simple composite mismatch analysis shows promise as an additional identification tool, although the data here are too sparse for making any major claims.

Discussion

Overall, Whole Word Match appears to be a simple measure that showed good agreement with researcher classification of children as TD/PPD in the crosslinguistic study. Out of 333 children, cut-off criteria for WWM accorded with research designation of participant status in 325 cases following the full word list and accepting degrooving as a match. The disagreements

included four false positives (German age 3, Canadian French age 3, Bulgarian, one at age 4 and one at age 5) and four false negatives (one Spanish age 3, and for European Portuguese, possibly one at age 3, and two at age 4). This represents a 97.5% hit rate, which is a commendable level for a simple measure that takes no qualitative information about the participant into account, and does not consider the type or proportion of mismatches. Mismatch analysis shows promise also as a potential additional tool in ambiguous cases (with far more data needed to make any substantive claims).

Overall, WWM appears to be a useful and efficient global measure for both research and clinical practice. It works in every language examined here, and there is even a basic similarity across languages in potential cut-off levels. All tests showed increasing cut-off levels for WWM across ages, enhancing construct validity.

Two languages showed lower cut-off scores than the others in the younger age groups (German, age 3 years; European Portuguese, ages 3 and 4 years) and one showed higher scores (Granada Spanish, ages 4 and 5 years). For German, the word list is comparable in complexity and length to the other languages, and thus was unlikely to be a factor, although the German test has a number of compound words, which may have negatively impacted scores in younger children (negative interactions of morphosyntax and phonology). For European Portuguese, the word set may be challenging for children under age 5 (150 total words, and over 47% of the words had three or more syllables, compared with fewer words and 25% multisyllabic words for the other languages). These more challenging tests (European Portuguese; possibly German) may be well-suited to older school-aged children and in any case, clinically, the criterion cut-off simply needs to be recognized as lower for those languages for preschoolers. The higher Spanish scores at age 4 and 5 years perhaps reflect the dialectal flexibility in Granada Spanish, where codas are optional, there are fewer clusters, and there are multiple acceptable variants for segmental production.

The Future of WWM: Bayesian Analysis

Toward the end of the WWM study, one of the authors of this chapter (Valter Ciocca) brought Bayesian analysis to the attention of the team as a possible statistical methodology that could reliably classify children on the WWM measure using a quantitative, probabilistic approach. For a test case, he used the Bulgarian 4-year-old data, where there was one child with PPD that appeared to be TD (the identity of whom he did not know in the data). The analysis was carried out on the full list (with degrooving counted as mismatch) using the R statistical software (R Core Team, 2018), and the Rethinking package for R (McElreath, 2016). Posterior distributions of probabilities assigned to WWM values were calculated using a beta-binomial model for group posterior, and a binomial model for individual posteriors. Figure 16 in the Appendix shows the posteriors calculated for individual children (PPD950, TD916, and PPD946) and the posterior for the whole group. The classification scheme is based on the overlap between group and individual posteriors, using the lower boundary of the 89% Highest Probability Density Interval (or HPDI) of the group posterior as the “cut-off point” (vertical dashed line in Figure 16). If more than 89% of the probability mass of an individual posterior is below the cut-off (to the left of the green line) then the child is classified as PPD (see PPD950, for whom 0.979 of probability mass is below the cut-off). If less than 89% but more than 50% of a child's posterior is below the cut-off, then the child is classified as "borderline” (see TD916;

probability mass below cut-off = 0.752). In all other cases, children are classified as "TD" (see PPD946; probability mass below cut-off = 0).

The Bayesian analysis concurred with the partially qualitative method used in the WWM study: PPD946 was classified as TD in both. The original researcher followed up on the child (now several years later) and determined that the apparent PPD at age 4 had resolved with no consequences; thus, the child may have been typical at the time of testing or PPD was identified for some reason other than WWM. Child performance is on a continuum, and forcing a binary division into PPD and TD may not be successful in all instances.

Implications of the Study for Research and Clinical Application

Whole Word Match and composite mismatch analysis are potentially reliable measures for classification of participants as TD/PPD, whether for research or clinical purposes. Clearly, larger sample sizes are needed for all languages, with younger and older children, and with a variety of dialects and languages. Bayesian data analysis appears to be a useful probabilistic method for identification of PPD on the basis of WWM data. The advantage of a classification scheme based on posterior distributions is that it is based on estimated uncertainty about who has PPD, and that the same classification criteria are widely applicable and independent of list-specific characteristics.

Clinical implications

For clinical practice, Whole Word Match has promise as a relatively quick measure for identification of PPD (and later, for possibly evaluating treatment outcomes). For phonological screening, an SLT could potentially learn to calculate WWM in an on-line task without phonetic transcription. With additional training, clinicians may in fact be able to apply this type of scoring metric in languages that they do not know, providing a measure for assessment and identification of PPD in those unfamiliar languages.

That the screening and full lists provide similar WWM data is also positive. The SLT can start with a screening set of words, and based on the preliminary cut-off scores as presented here plus other important information about the client, decide whether to continue with the full set of words or to stop testing. If the full test does not clearly identify PPD, a composite mismatch analysis may further elucidate classification, because a 0-1 match score says nothing about the number and types of mismatches that result in a "0" WWM score for a word. Furthermore, as noted earlier, other factors about the client also pertain to identification of PPD (e.g., oral mechanism, hearing, cognition, general language abilities, perception, environmental context, social needs, literacy). Should intervention be indicated, we emphasize further that treatment planning will require phonetic transcription and phonological analysis of the full list to maximize the opportunities for successful outcomes.

As an additional note, for preschoolers, it appears that many people consider degrooved sibilants to be typical in acquisition, because of the high proportion of children designated as TD who showed this pattern across the languages. Thus, degrooving is not necessarily a contributing factor to PPD in the preschool years.

Finally, we encourage the readers to use the materials and tutorials on our free website, phonodevelopment.sites.olt.ubc.ca. As materials come available, including criterion reference data for a number of measures, including WWM and mismatch levels, we will add these to the

website. We also encourage others to consider contribution both to that website, and Phonbank (www.phon.ca), in the ever-growing database of information about children's speech development.

Acknowledgments

The authors acknowledge the children, their parents and the many research assistants and volunteers who have made this research possible. We are also indebted to the Social Sciences and Humanities Research Council for funding (410-2009-0348, 611-2012-0164) and to agencies in the various countries for their support.

Declaration of interest

The authors have no conflict of interest with any of the material in this report.

References

Arias, J., & Lleó, C. (2013). Rethinking assessment measures of phonological development and their application in bilingual acquisition. *Clinical Linguistics and Phonetics, 28,* 153–175.

Bernhardt, B. H., & Stemberger, J. P. (1998). *Handbook of phonological development: From a nonlinear constraints-based perspective*. San Diego: Academic Press, now with Emerald Group Publishing.

Bernhardt, B. M., & Stemberger, J. P. (2017). Investigating typical and protracted phonological development across languages. In E. Babatsouli, D. Ingram & N. Müller (eds.) *Crosslinguistic encounters in language acquisition: typical and atypical development* (pp. 71-108). Bristol, UK: Multilingual Matters.

Bernhardt, B. M., & Stemberger, J. P. (2012). Translation to practice: Transcription of the speech of multilingual children. In B. Goldstein & S. McLeod (eds.) *Multilingual Aspects of Speech Sound Disorders in Children* (pp.182-190). Bristol, UK. Multilingual Matters.

Edwards, M. L. (1992). Clinical forum: Phonological assessment and treatment in support of phonological processes. *Language, Speech, and Hearing Services in Schools, 23*, 233–240.

Hedlund, G., & Rose, Y. (2018). *Phon 3.0.2* [Computer Software]. Retrieved from
https://phon.ca December 23, 2018.
Language, Speech, and Hearing Services in Schools, 14, 210–214.
http://www2.waisman.wisc.edu/phonology/pubs-tech.html. Retrieved December 23, 2018.

Ingram, D. (2002). The measurement of whole-word productions. *Journal of Child Language, 29,* 713–733.

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.

Mason, G. (2018 online). School-aged children's phonological accuracy in multisyllabic words on a whole-word metric. *Journal of Speech, Language, and Hearing Research*, 1-15. doi:10.1044/2018_JSLHR-S-17-0137.

McElreath, R. (2016). *Rethinking package for R*, version 1.59. (https://github.com/rmcelreath/rethinking)

Preston, J. L., Ramsdell, H. L., Oller, D. K., Edwards, M. L., & Tobin, S. J. (2011). Developing a weighted measure of speech sound accuracy. *Journal of Speech, Language, and Hearing Research, 54*, 1–18.

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria (URL http://www.R-project.org/).

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.* 308-401*)*. Oxford, UK: Oxford University Press.

Schmitt, L. S., Howard, B. H., & Schmitt, J. F. (1983). Conversational speech sampling in the assessment of articulation proficiency.

Shriberg, L. D., Austin D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The Speech Disorders Classification System (SCDS): Extensions and lifespan reference data. *Journal of Speech, Language and Hearing Research, 40*, 723–740.

Van Borsel, J. & D'haeseleer, L. (2018 online). The Process Density Index as a measure of phonological development: Data from Dutch. *Communication Disorders Quarterly,* https://doi.org/10.1177/1525740118790532.

Appendix (FIGURE CAPTIONS AND 16 FIGURES ARE IN A SEPARATE FOLDER).

Table 1. Numbers of words on the phonological evaluations and participants by age.

Language Family	Language	Number of test words[a]		Age 3		Age 4		Age 5	
		Screener	Full List	TD	PPD	TD	PPD	TD	PPD
Germanic	German (Cologne)		105	9[b]	9[b]	10[b]	10[b]		
	Icelandic	39 (50)[c]	84 (110)[c]	14[b]	14[b]	14[b]	13[b]		
	Swedish (Linköping)	35	109			12[b]	12[b]		
Romance	Canadian French	46	111	6	7	7	2		
	Portuguese (Lisbon)	50	150	27	1	30	1		
	Spanish (Granada)	39 (40)[d]	103	10	8	9	13	11	8
South Slavic	Bulgarian (Sofia)	50	111	10	10	10	10	10	10
	Slovenian (Central)	48 (50)[d]	101			8[b]	8[b]		

[a] On the full tests, 5-10% of words were repeated for most children to evaluate consistency and increase the number of tokens for infrequent phonemes. Here we give only the number of different words.
[b] Exact age- (and gender-) matched controls.
[c] Icelandic: Tests for PPD/TD were not identical: analysis considered only words in common; there were no significant differences for Whole Word Match between all words tested and the set compared.
[d] Tests have been slightly revised since original data collection. The new screeners have 1-2 words that are not in the words analyzed here.

Table 2. Preliminary Percent Whole Word Match criterion levels for identification of protracted phonological development (PPD).

Language	Age 3		Age 4		Age 5	
	Screener	Full list	Screener	Full list	Screener	Full list
German		30		50		
Icelandic	40	40	50	60		
Swedish			50	50		
Canadian French	45	50	50	55		
European Portuguese	15	25	40	40		
Granada Spanish	40	40	65	60	80	75
Bulgarian	40	40	60	60	55	55
Slovenian			60	60		

Note. These cut-off criteria are based on very small samples and count all mismatches, including degrooved sibilants (Except in Swedish where such were ignored). Levels would be slightly higher if degrooved sibilants were considered a match. Larger groups are needed to establish norm references. For individual children, other factors always must be taken into account during the process of identification of PPD (intelligibility, social needs, oral mechanism and hearing factors, cognition, literacy).

Table 3. Example of disambiguation of identification through a composite mismatch proportion analysis: Spanish

Original classification	Age (mo.)	WWM Full List	Mismatches/ total words	Proportion of mismatch types/words CDel	CSub	V changes	Epen	Clinical decision
TD326	37	39.4	**.93**	**.19**	**.56**	.11	.08	Review in 1 year?
PPD307	38	38.2	**1.06**	**.29**	**.46**	**.18**	**.13**	Review in 3 months? Treat?
PPD330	39	34.9	**1.15**	.27	**.84**	.02	.02	Review in 3 months? Treat?

Note. Vowel changes = vowel (syllable deletion), diphthong reduction or substitutions. CDel = Consonant deletion; CSub = consonant substitution; Epen = epenthesis; TD = typically developing, PPD = protracted phonological development.