

## RESEARCH REPORT

# Factors affecting judgment accuracy when scoring children's responses to non-word repetition stimuli in real time

Peter Howell<sup>1</sup>  | Clarissa Sorger<sup>1</sup> | Roa'a Alsulaiman<sup>1,2</sup>  | Kaho Yoshikawa<sup>1</sup> | John Harris<sup>1</sup>  | Kevin Tang<sup>1,3,4</sup> 

<sup>1</sup>Division of Psychology and Language Sciences, University College London, London, UK

<sup>2</sup>Department of Psychology, Education College, King Saud University, Riyadh, Saudi Arabia

<sup>3</sup>Department of Linguistics, University of Florida, Gainesville, Florida, USA

<sup>4</sup>Department of English Language and Linguistics, Institute of English and American Studies, Heinrich-Heine-University, Düsseldorf, Germany

## Correspondence

Peter Howell, Department of Experimental Psychology, University College London, Gower Street, London WC1E 6BT, UK.

Email: [p.howell@ucl.ac.uk](mailto:p.howell@ucl.ac.uk)

Kevin Tang, Department of English Language and Linguistics, Institute of English and American Studies, Heinrich-Heine-University, Düsseldorf DE-40225, Germany.

Email: [kevin.tang@hhu.de](mailto:kevin.tang@hhu.de)

## Abstract

**Background:** Non-word repetition (NWR) tests are an important way speech and language therapists (SaLTs) assess language development. NWR tests are often scored whilst participants make their responses (i.e., in real time) in clinical and research reports (documented here via a secondary analysis of a published systematic review).

**Aims:** The main aim was to determine the extent to which real-time coding of NWR stimuli at the whole-item level (as correct/incorrect) was predicted by models that had varying levels of detail provided from phonemic transcriptions using several linear mixed method (LMM) models.

**Methods & Procedures:** Live scores and recordings of responses on the universal non-word repetition (UNWR) test were available for 146 children aged between 3 and 6 years where the sample included all children starting in five UK schools in one year or two consecutive years. Transcriptions were made of responses to two-syllable NWR stimuli for all children and these were checked for reliability within and between transcribers. Signal detection analysis showed that consonants were missed when judgments were made live. Statistical comparisons of the discrepancies between target stimuli and transcriptions of children's responses were then made and these were regressed against live score accuracy. Six LMM models (three normalized: 1a, 2a, 3a; and three non-normalized: 1b, 2b, 3b) were examined to identify which model(s) best captured the data variance. Errors on consonants for live scores were determined by comparison with the transcriptions in the following ways (the dependent variables for each pair of models): (1) consonants alone; (2) substitutions, deletions and insertions of consonants identified after automatic alignment of

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *International Journal of Language & Communication Disorders* published by John Wiley & Sons Ltd on behalf of Royal College of Speech and Language Therapists.



live and transcribed materials; and (3) as with (2) but where substitutions were coded further as place, manner and voicing errors.

**Outcomes & Results:** The normalized model that coded consonants in non-words as ‘incorrect’ at the level of substitutions, deletions and insertions (2b) provided the best fit to the real-time coding responses in terms of marginal  $R^2$ , Akaike’s information criterion (AIC) and Bayesian information criterion (BIC) statistics.

**Conclusions & Implications:** Errors that occur on consonants when non-word stimuli are scored in real time are characterized solely by the substitution, deletion and insertion measure. It is important to know that such errors arise when real-time judgments are made because NWR tasks are used to assess and diagnose several cognitive–linguistic impairments. One broader implication of the results is that future work could automate the analysis procedures to provide the required information objectively and quickly without having to transcribe data.

#### KEYWORDS

bilingualism, diversity, English as an additional language, non-word repetition, universal non-word repetition

#### WHAT THIS PAPER ADDS

##### *What is already known on this subject*

- Children and patients with a wide range of cognitive and language difficulties are less accurate relative to controls when they attempt to repeat non-words. Responses to non-words are often scored as correct or incorrect at the time the test is conducted. Limited assessments of this scoring procedure have been conducted to date.

##### *What this study adds to the existing knowledge*

- Live NWR scores made by 146 children were available and the accuracy of these judgements was assessed here against ones based on phonemic transcriptions. Signal detection analyses showed that live scoring missed consonant errors in children’s responses. Further analyses, using linear mixed effect models, showed that live judgments led to consonant substitution, deletion and insertion errors.

##### *What are the practical and clinical implications of this work?*

- Improved and practicable NWR scoring procedures are required to provide SaLTs with better indications about children’s language development (typical and atypical) and for clinical assessments of older people. The procedures currently used miss substitutions, deletions and insertions. Hence, procedures are required that provide the information currently only available when materials are transcribed manually. The possibility of training automatic speech recognizers to provide this level of detail is raised.

## INTRODUCTION

Non-word repetition (NWR) tests are used for assessing language development and a wide variety of psychological conditions (Archibald, 2008). During NWR testing, participants hear and imitate phone strings that are not words in the language they speak. Poor NWR performance indicates that cognitive processing is affected and can help identify whether the phonological component of working memory contributes to any impairment that arises (Gathercole et al., 1994). A spoken response is always required in NWR tests.

Different procedures for stimulus presentation and analysis of responses are permitted during NWR testing. To establish what variant forms are used, a secondary analysis of an existing systematic review (Schwob et al., 2021) is reported in Experiment 1. This showed that one important difference between procedures was whether results were obtained live as the participant made a response or if materials were recorded and transcribed and analysed later offline. The merit in using transcription-based analysis procedures is the detail provided, but adopting them is usually impractical in clinics and schools. Although transcription may be used rarely when data are collected in these settings, it is still important to document what information about NWR is missed when responses are scored live. To address this, Experiment 2 compared live and transcription-based scoring of the same materials.

## EXPERIMENT 1

### Introduction

Documentation about procedures used across NWR studies, such as how stimuli are played to participants and how often participants' responses are scored in real time (live), is not currently available. Whilst rigorous procedures should be adopted where possible, practical considerations do not always allow for them to be employed in all testing environments. Whilst this situation obtains: (1) clear statements about what procedures were used in studies is necessary (Experiment 1); and (2) information is required concerning what impact the several different scoring procedures have on the accuracy of results (Experiment 2 starts to address this question).

To address the first issue, a re-analysis of the studies identified in Schwob et al.'s (2021) recent systematic review was conducted to obtain further information about procedures used in studies. Schwob et al.'s review addressed how well NWR tests identified developmental language

disorder (DLD) for studies that used samples of children with heterogeneous language backgrounds. Schwob et al. (2021) was used as a starting point because studies included in this review met rigorous standards for report of procedures, and their work, and our own, have a shared focus on using NWR as a diagnostic tool to identify DLD. Moreover, the studies in Schwob et al.'s review used children with similar ages (up to the age of 8;11) to those examined in Experiment 2. Schwob et al.'s evaluation of scoring procedures was limited to reports of whether the percentage of items correct (whole item) or percentage of phonemes correct were reported, whereas our focus was on comparing real time versus offline transcription-based scoring procedures.

The re-analysis required details to be reported of factors that were not evaluated by Schwob et al. (2021). This resulted in some studies that passed Schwob et al.'s quality criteria being excluded if they did not report on the extra details required here. The first additional detail was to document whether real-time or recorded material was used in analyses. This is necessary because real-time assessments should reduce accuracy (increase misses and/or lead to false alarms) more than assessments based on transcriptions. Holistic judgements are always made when material is judged live and can also be made later when recordings are scored. Although holistic judgments can be made on live and transcribed responses, different NWR scores would arise because of the different judgment demands in the two contexts. Reliable information about phonemic errors is only possible (but not always obtained) when recordings are available for offline analysis after data collection.

The second detail was whether the NWR stimuli were spoken live by the experimenter or if pre-recorded material was used. Using recordings ensures that test stimuli are constant across the children tested. Whilst recording children's responses makes some equipment demands and is, arguably, not necessary if detailed transcription analysis is not conducted, this justification applies to a lesser extent to stimulus materials (e.g., a mobile phone is convenient for playing pre-recorded stimuli). To this end, high-quality recordings are available online for some NWR tests such as Gathercole et al.'s (1994) popular child non-word repetition (CNRep) and Howell et al.'s (2017) universal non-word repetition (UNWR) tests (<https://www.fistproject.org/resources/>).

Finally, it was of interest whether different analysis procedures were favoured across countries. In particular, do different countries prefer real time versus transcription-based procedures?

## Aims

The studies included in Schwob et al. (2021) were examined to document the types of procedures used in NWR work. Information about the following details was obtained:

1. Across studies, how frequently are NWR responses assessed in real time? As a follow-up to this issue, it was also determined for the studies that used recorded responses in analyses, whether analysis was whole-word or phoneme-based (similar to what Schwob et al. reported for the entire cohort of studies).
2. The incidence with which NWR stimuli were spoken live to participants. A secondary question about stimuli was what presentation format was used (audio or other, usually audiovisual—AV).
3. Whether analysis procedure adopted in studies depended on the country where the study was conducted?

## Method

Schwob et al. (2021) identified 46 studies that passed their quality criteria for inclusion in the review. A total of 45 of these studies were available here (the one omitted was published in an expensive book that was not available). The methods used in the 45 studies were examined in the first pass to extract the following information by author RA who arranged this information in a 45-row table with columns as follows:

1. Response scoring: Were materials scored in real time or were they based on offline transcription-based analyses using recordings? For studies analysed offline, were analyses holistic or phoneme based?
2. Stimulus presentation: Did participants hear pre-recorded items or were the stimuli produced by the experimenter? Also, what was the mode of stimulus delivery?
3. The country in which the research took place was determined from the affiliation of the first author.
4. Author PH checked all information in the table. At this stage, a further study was dropped because it lacked many details (44 studies left in the analyses). RA and PH then conferred and resolved any discrepancies in their designations about each study.

## Results

Primary data for the first two details (whether responses were scored from recordings and mode of stimulus

delivery) are presented in Table 1. Stimulus delivery options (top axis) were pre-recorded, live or not clear. Data formats for response scoring (side axis) were recorded, not recorded and not clear. Overall, 32 studies used pre-recorded stimuli (left-hand column of Table 1). For these, the material used for analysis was based on recordings, live responses or was unclear for 22, seven and three studies. Nine studies used real-time spoken stimuli (middle column of Table 1) and responses were recorded, not recorded or unclear in four, three and two studies, respectively. Finally, there were three studies where it was not possible to identify how stimuli were presented, where there was one each where response analysis was based on recordings, online or indeterminate (column 3 of Table 1).

The 22 studies that used recorded stimuli and took response recordings for offline analysis (top left entry of Table 1) were examined to determine whether whole- or part-item scoring procedures were used. The four studies that did not use pre-recorded stimuli but did analyse recorded responses were excluded from this analysis because response scoring for them depends on veridicality and stability over test occasions of the stimuli presented which cannot be guaranteed.

All the 22 remaining studies used part-word scoring procedures and audio-formatted data in analyses. The scoring procedures varied markedly across studies but are not described since they are beyond the scope of the current work. The mode of stimulus delivery for these 22 studies was predominantly audio (18/22, 81.8%).

The entire cohort of 44 studies included 16 from the United States and was five or fewer for other countries (e.g., only two were from the UK). The majority of the US studies used pre-recorded stimuli and recorded responses (10/16 = 62.5%), none used live stimuli and recorded responses, four used pre-recorded stimuli but did not record responses (25%), and two used live stimuli and did not record responses (12.5%). The high rate of use of pre-recorded stimuli and responses in US studies (62.5%) exceeded the rate for other countries (51.7%). Numbers of studies was too low for analysis for other countries.

## Discussion

In summary, a third of the studies included in the systematic review did not take recordings and this necessitated scoring children's responses live; the country that provided most of these studies (17/45, 37.7%) was the United States with only small numbers from other countries. For those studies where recordings were available, all used audio records alone for analyses and audio stimulus presentation predominated (81.8%). Only 50% of the studies that used recordings for determining responses (a third of all 44

**TABLE 1** Number of studies meeting the contingencies at the top and side from the re-analysis of the papers in Schwob et al.'s (2021) systematic review.

		Stimulus presentation		
		Pre-recorded, 32	Live, 9	Not clear, 3
Response recorded	Yes	22	4	1
	No	7	3	1
	Not clear	3	2	1

Note: The top row indicates how stimuli were presented (pre-recorded, live or study was unclear). The side row indicates whether or not recorded responses were obtained or if this was unclear.

studies) used pre-recorded stimuli. All of these 50% of studies that used recordings analysed parts of NWR stimuli.

Several studies cross-referred to Dollaghan and Campbell (1998) and indicated that they followed their procedures rather than included details in reports. In Dollaghan and Campbell's study: (1) non-words were pre-recorded. They were spoken by a trained female speaker who had practiced producing each non-word at a consistent rate; (2) non-words were presented over headphones in a quiet location using a good-quality cassette recorder; (3) responses were audio recorded by an external microphone for phonetic transcription; (4) scoring was part-word as each phoneme (consonant or vowel) was scored in relation to its target phoneme; and (5) details were given on what were considered phoneme errors. Clearly, the early Dollaghan and Campbell study met and surpassed stimulus delivery and response analysis procedures required for inclusion in the current analysis. Studies citing Dollaghan and Campbell may not have followed all these criteria, but it was not possible to decide this from the information provided. It seems likely that aspects (3) and (4) were usually followed, but it is less clear about the other aspects. Nevertheless, all these procedural details were presumed to be met. However, future work should be specific whether some or all of Dollaghan and Campbell's procedures were followed. For example, it seems unlikely that cassette recorders (2) and external microphones (3) are used in contemporary studies.

It was originally planned to determine how often speakers were visible when stimulus mode was AV and whether and how AV information was used during response scoring. A small number of studies used AV as stimuli (six in total) and these predominantly scored children's responses in real time (4/6). In these cases, children saw the experimenter speaking and the experimenter saw the children responding but details about whether and how the experimenter assessed children's AV responses to the NWR material were lacking. For instance, details were not given about whether children focused on the speaker's articulators nor even whether children looked at the speaker. An additional important caveat about use of AV presentation applies. Whilst a real-time AV format for stimulus presentation and assessment of children's responses is desirable

for SaLTs and could, in principle, be made in all settings even when audio alone is used in analyses, video recordings cannot be obtained for response recording in schools to ensure that children cannot be identified (Dockrell & Howell, 2015). The mismatch between use of audio and video response recordings being available in clinics versus audio alone when recordings need to meet anonymity requirements limits what comparison can be made across studies using different record formats.

It proved difficult to determine whether holistic responses or phonemic accuracy was scored in several studies. The main reason was that reports of phonemic scoring were often underspecified as noted in connection with the subset of studies that cited Dollaghan and Campbell (1998). Some flexibility was allowed in analysis procedures here insofar as they did not need to follow a published NWR test protocol when one was available. This allowed for studies to make slight variations on the stimulus presentation/item scoring from those published depending on the aim of the paper. For example, some studies used published test items but scored response differently to that specified in the publication. Whilst it may legitimately be assumed that live presentations were judged holistically, what was done with recordings was unclear. When phone errors were scored, transcription is implied and any comments concerning the error types that ought to make an item warrant a response of incorrect should be noted (see Experiment 2 for the alternative error taxonomies used in that work). However, clear statements were not available in the majority of reports. The main question that arises from this study concerns the accuracy of the materials scored live (a third of all the studies considered). This is addressed in the following experiment.

## EXPERIMENT 2

### Introduction

Experiment 1 established the need to improve documentation of procedures used when scoring NWR materials and to then investigate the impact of using different scoring procedures (real-time and various schemes that can

be applied with transcriptions) on NWR results. As a step toward the latter, Experiment 2 compared NWR holistic responses made live (Howell et al., 2017) against phoneme-based error scores that used different error-taxonomies applied to the same response materials. As background, the following provides: (1) a description of some of the features of the NWR task employed (specifically, how it applies to children from diverse language backgrounds); (2) consideration of what factors are important to incorporate into transcription-based scoring procedures for assessing typical and atypical speech development; (2a) argues for using measures of substitutions, deletions and insertions in taxonomies. The issue is also raised about whether breaking substitutions down further by the phonemic properties of place, manner and voicing could provide a useful additional parameter to characterize NWR performance of children's language performance; (2b) describes how substitutions, deletions and insertions can be computed. The factors and methodological approaches considered in 2a and 2b are the basis on which models to compare live versus transcription data were developed and applied in this study; and (3) the aims of the present study are summarized.

### **NWR testing for samples of children who use diverse languages**

NWR tests face new challenges internationally. For instance, Howell et al.'s (2017) UNWR test was developed for assessing children's speech when they start at UK schools. Two problems it had to address were that many children did not use English in the home and a wide variety of alternative languages was used. By definition, NWR tests do not include word material for the target language. However, if the NWR test is to be used with speakers of alternative languages, language-specific biases that make some stimuli word-like for any of these languages other than the target language can arise. Empirical studies show that these biases affect results. Thus, NWR tests designed for one language lead to superior performance for children who use that language compared with children who use another one (Masoura & Gathercole, 1999; Windsor et al., 2010). This would lead to variation in performance due to language that the child uses unless the biases are controlled.

### **How the design of UNWR allows assessment of participants from diverse language groups**

Howell et al.'s (2017) UNWR test provides an unbiased assessment that applies to all the languages that children in their test cohort spoke (children starting in UK schools).

UNWR has the following rules that generate non-word materials that apply to a set of 20 or more languages: (1) UNWR uses a fixed set of consonants that are used in all the languages in the set; (2) the phonotactic properties of onsets and codas for individual syllables that also apply across the set of languages were identified—this allowed NWR materials to be generated that extended to structures additional to single consonant–vowel syllables (CVs) that permit non-words with onset and coda strings; (3) phonotactic rules for concatenating syllables that apply across all target languages were specified which allowed multi-syllabic test material to be generated; and (4) exemplars consisting of all strings meeting the previous constraints were generated automatically.

The materials that were generated according to these rules were potential candidates for NWR stimuli, but needed checking for word-likeness across the languages. Where there were existing dictionaries for any of the languages, these were used to exclude word forms (Howell et al., 2017). When dictionaries were lacking, these were generated and similar word-like exclusions were applied.

NWR tests with the UNWR materials were made for the children whose data were used in the current study. The performance of these children assessed in real time was reported in Howell et al. (2017) and these provided the live responses for comparison with transcribed materials here. Vowels were ignored during scoring in Howell et al. because few vowels are used consistently across languages. This is supported by Chiat (2015) whose Crosslinguistic nonword repetition (CLT) test only used three vowel qualities (/a, i, u/) since these are the only ones used commonly across languages. This is also suggested by how the functional load of vowels is generally higher than that of consonants cross-linguistically according to corpus analyses (Oh et al., 2013, 2015) and by demonstrations that reducing vowels to schwa in synthetic sentences has minor effects on intelligibility (Whiteside, 1996).

### **Scoring procedure for assessing typical and atypical speech development based on transcriptions**

#### **Factors to score to provide information about typical/atypical language development**

NWR tasks were first introduced as a way of assessing cognitive functions in children. Consequently, analysis procedures need to be tailored if they are to provide useful information specifically about speech and language development. In particular, ideally scoring parameters ought to address issues with phonemic and phonological processes that some children experience. These parameters need to capture dynamic age-related changes to speech because

**TABLE 2** Dodd's (2005) phonological processes associated with delayed (T) or atypical (A) language development with examples on each process.

T/A process	Example	S/I/D	P/M/V
T Labial assimilation	'mad' as 'mab'	S	Place
T Velar assimilation	'take' as 'kake'	S	Place
T Alveolar assimilation	'time' as 'tine'	S	Place
T Prevocalic voicing	'pig' as 'big'	S	Place
T Stopping	'zoo' as 'do'	S	Manner
T Fronting	'cat' as 'tat'	S	Manner
T Deaffrication	'chip' as 'ship'	S	Manner
T Gliding	'rabbit' as 'wabbit'	S	Manner
T Vowel shortening	'third' as 'thud'	S	Manner
T Devoicing	'bad' as 'bat'	S	Voicing
T Reduplication	'daddy' as 'dada'	S	Any
T Final consonant deletion	'cat' as 'ca'	D	-
T Cluster reduction	'play' as 'pay'	D	-
A Backing	'time' as 'kime'	S	Place
A Stop replaces glide	'yes' as 'des'	S	Manner
A Fricative replaces stop	'sit' as 'sis'	S	Manner
A Glottal replacement	k in 'pick' as glottal stop	S	Manner
A Initial consonant deletion	'cut' as 'ut'	D	-

Note: Whether the processes involved substitution (S), insertion (I) or deletion (D) (column 3) and place (P), manner (M) and voicing (V) (column 4) are given.

phone and phonological process issues can be delayed, requiring determination whether language development is slow but not disordered. However, comprehensive information about age-of-acquisition of phones in typical and atypical speech are not available to incorporate into scoring procedures.

As an interim solution, word usage metrics can be used to incorporate contributions from phonological factors and, furthermore, substitutions, which are a type of usage error, can be broken down into place, manner and voicing errors to provide some information on difficulties on particular phonemes. The link between conventional phonological processes and the usage parameters is seen by considering Dodd's (2005) work that distinguished typical from atypical phonological processes. Usage metrics identify three aspects: substitutions (e.g., 'bat' for 'pat'), deletions ('smell' for 'smelt') and insertions (e.g., 'going' for 'gone'). Typical processes happen frequently up to certain ages for many if not all children, but some children persist in using them beyond these ages and they then have delayed phonological development. Other children use more idiosyncratic phonological processes and their use may indicate atypical phonological development. Table 2 lists the processes Dodd associated with either delayed or atypical language development. The substitution and deletion parameters characterize all items on this list as indicated in column 3.

Substitutions are broken down with respect to place, manner and voicing changes on the substituted phones (right-hand column of Table 2) to provide partial indications of difficulties with individual phoneme types. There are no examples of insertion in Table 2, but this can be a characteristic of delayed and/or atypical speech development. For example, consonant epenthesis ('toptic' for 'topic' or 'play' for 'pay') is a characteristic of typical speech development and glottal stops or /h/-insertions are the most common consonant epentheses.

Thus, there is a link between phonological processes that are monitored for determining delayed and atypical phonological processing in children, on the one hand, and the usage parameters of substitution, deletion and insertion and the phonemic parameters of place, manner and voicing in substitutions on the other. This suggests that using the latter parameters in analyses would assay issues with phonological and phonemic processes that children may have without waiting for comprehensive documentation of phonological and phonemic issues to become available. In sum, the advantage in using transcriptions and coding them with the usage and phonemic parameters indicated is that they incorporate factors that mediate aspects of typical and atypical phonological development in a way that is practicable at present.

## Computation of substitutions, insertions and deletions: Levenshtein distance metrics

The analysis approach adopted here was to fit three models that compared real-time judgments and transcriptions of the Howell et al. (2017) dataset. Usage factors were considered a proxy for representing issues in phonological development and the substitutions were subsequently divided into place/manner/voicing errors. The way that the usage factors were scored was based on the neighbourhood density approach which compares materials (here, transcriptions and targets) by identifying phones that have been substituted, deleted or inserted in one utterance (e.g., a child's response) relative to another (e.g., a transcribed response or a non-word stimulus heard). To illustrate with the strings /'blɪmpɪək/ and /'blɪpɪək/. The only difference is that /m/ is deleted in the second string.

Levenshtein distance metrics can be used to indicate how the NWRs differed from the target pronunciation of each non-word in terms of substitutions, deletions and insertions. The non-word transcriptions of the attempts can be segmentally aligned with the corresponding target non-word transcriptions (the target stimuli presented). An automatic alignment method based on the Pointwise-Mutual-Information-based Levenshtein distance (Wieling et al., 2009), as adapted by Tang (2015: ch. 2), can be used to estimate Levenshtein's distance. To illustrate how this works, if [S K AE T] (Target) was to be aligned with [P AE T] (Attempt), the best alignment would align [S] with nothing (i.e., it is deleted), [K] with [P], [AE] with [AE], and [T] with [T]. This alignment relies on the intuition that [K] is phonetically more similar to [P], than [S] is to [P]. The advantage of this automatic alignment method is that such phonemic knowledge is automatically derived from the alignments themselves, particularly for cases with straightforward alignments, for example, as when [K AE T] (Target) was aligned with [P AE T] (Attempt).

These straightforward alignments from mispronounced non-words form the basis of the phonetic relationship between phonemes. If a phoneme is frequently mispronounced as another phoneme, then they are likely to be phonetically similar. The alignment algorithm contains parameters that can influence the preferences for aligning specific segments with other specific segments, for instance, having a preference for aligning vowels with vowels, and consonants with consonants and a dis-preference for aligning vowels with consonants and vice versa. A cost value is specified for each of the possible combinations of segments, the higher the cost value, the higher the dis-preference the algorithm would have for aligning the corresponding combination of segments. Using this method, all pairs of non-words would first be aligned with a general dis-preference for aligning a consonant (excluding

glides /j/ and /w/) with a vowel and vice versa. Such alignments are achieved by assigning all consonant–vowel and vowel–consonant pairs with an arbitrarily higher cost than all other combinations. The output of the first round of alignment is a set of aligned phone pairs. These aligned phone pairs are then used as the basis of the cost of a phone aligning with a different phone for the next round of alignments. The costs were assessed by, first, counting the frequencies of a given aligned pair (Phone<sub>A</sub> and Phone<sub>B</sub>), and the two phones; second, these frequencies were used to compute the probability of aligning Phone<sub>A</sub> and Phone<sub>B</sub>, and the probabilities of Phone<sub>A</sub> and Phone<sub>B</sub>; third, a Pointwise Mutual Information score was estimated which compared the probability of observing Phone<sub>A</sub> with Phone<sub>B</sub> with the probability of observing these two phones independently (i.e., chance); and, finally, converting the score to a cost value for Phone<sub>A</sub> and Phone<sub>B</sub>. A higher Pointwise Mutual Information score indicates the two phones are more likely to co-occur, and the more likely it is that two phones co-occur, the lower the cost is for that phone pair. The alignment procedure is repeated as many times as required until there is no change in the alignments. This method therefore minimizes potential influences of relying on a given phonetic/phonological theory of segmental relationship. The resultant alignments are then used to identify which segments in the Target were substituted deleted or inserted with those in the Attempt as determined from the transcription of that Target.

The alignment procedure is illustrated with an example of an actual misproduction. The Target [B R AE F R IX] was produced as [G L AE F AX L IX] and the alignment was [B:G R:L AE:AE F:F -:AX R:L IX:IX] where ':' denotes 'aligned with' and '-' denotes an empty phone (the alignment is shown in Figure 1). Crucially, for the second syllable the algorithm sensibly aligned [R] with [L] and [-] with [AX] which indicates a vowel insertion, as opposed to [R] with [AX], and [-] with [L]. The alignment is 'sensible' since misproducing [R] as [L] and inserting a schwa is more likely than misproducing [R] as a schwa and inserting [L].

## Normalized/non-normalized Levenshtein distance metrics when incorporating usage metrics and phonemic properties for scoring NWR

Normalization is motivated by the fact that longer words are likely to have a higher distance score, therefore dividing the distance by the alignment length controls for that property. The use of distance is to approximate human perception of the target word and the mispronounced word. However, it is unclear whether listeners perceive



B	R	AE	F	-	R	IX
G	L	AE	F	AX	L	IX

An example of a good alignment

B	R	AE	F	R	-	IX
G	L	AE	F	AX	L	IX

An example of a bad alignment

FIGURE 1 Examples of good (top) and poor (bottom) alignments after calculation of the estimated Levenshtein's distance.

differences in terms of the absolute number of differences (non-normalized distance) or in terms of the proportion of a word being different (normalized distance) as shown by the following studies.

Levenshtein distance is used in dialectology research to compute the distance between pairs of cognates. While some studies (Schepens et al, 2012) employed normalized distance, others (Heeringa et al., 2006) report that non-normalized distances are better at approximating dialect differences as perceived by the dialect speakers than are normalized distances. Also, Bailey and Hahn (2001) compared different phonotactic probability measures and reported that a non-normalized measure of phonotactic probability (which penalizes longer words more harshly than shorter words) provided a modest, but consistent, gain in variance explained in a word-likeness judgment task. Given these findings, best practices about whether or not to normalize by length remain to be established (Nerbonne et al., 1999). For these reasons, both normalized distance and non-normalized distance were examined in this study.

## Aims

As noted, transcriptional analysis is usually not feasible. Nevertheless, understanding what types of errors can be missed in real-time judgments is important: (1) to show what type of phonological processes are being underestimated/failing to be examined; (2) to inform practitioners about how detailed their transcription needs to be for different purposes; and (3) to provide training/specific instructions for live scorers to pay specific attention to things that they are likely to miss.

To start to address these issues, Experiment 2 attempted to determine what types of error are missed when live scores are made and compared with transcriptions. The 'live' scored UNWR responses from Howell et al. (2017) were compared with those from phonemic transcriptions of the same material made offline. Each live score provided a real-time binary indication about whether the

consonants in the non-word were judged to have been spoken correctly or incorrectly. The corresponding transcriptions provide a basis for determining what factors affected accuracy of the live score. Signal detection analysis was conducted first that dissociates two forms of error: (1) misses where, in the present study, real-time judgments fail to identify individual consonants that appear in the transcriptions and (2) false alarms where real-time judgements identify a consonant that is absent in the transcriptions. The demands involved in real-time judgments predict that judgement errors are more likely to lead to misses than to false alarms.

Next, three linear mixed models (LMMs) were fitted to the data to establish sources of discrepancy between live judgments and the transcriptions. The fixed effect predictors of these discrepancies in each model were: (1) consonants alone: This model used composite consonant error scores as Levenshtein distances across all consonants irrespective of error type (substitution, deletion or insertion) and across consonant features (place, manner, voicing); and (2) substitution, deletion and insertion parameters derived from usage work: Here the model incorporated Levenshtein distances for substitutions, deletions or insertions of consonants in the response transcriptions as compared with the standard (stimulus) forms (Heeringa, 2004); (3) as in (2) but with substitutions divided into those involving place, manner or voicing errors where each type of error was quantified in terms of its Levenshtein distance.

Models with (1a, 2a, 3a) and without (1b, 2b, 3b) normalization were fitted. The same set of random control variates was included in all models. The control variates were *vowel* Levenshtein distance, number of target consonants, trial position and bilingualism. These were included to determine whether they affected model fit (no effects were expected). For instance, the judges who made the live scores were instructed to ignore vowels and focus on whether the consonants were correct (identity and position). Thus, it would be expected that performance would vary with the number of consonants, but not vowels, in a non-word stimulus.



The predictions were as follows: model 1 (the baseline) should produce poorest fit to the data because it does not include any information related to phonological or phonemic processes that can affect children's responses. There are reasons both why model two should provide a better fit than model three and vice versa. Model two includes substitutions and this generic measure may capture more about child language behaviour and the errors that arise when scoring them than breaking them down into place, manner and substitution components. On the other hand, information about specific phoneme difficulties (which the division of types of substitution provides) may enhance model fit. Hence, no directional predictions were made about whether model two or model three should provide best fit to the data. However, predictions were made about the order of importance of place/manner/voicing confusions for model 3. Crowe and McLeod (2020) reviewed articles concerning consonant age of acquisition in English-speaking children in the United States, and reported that plosives, place and voicing were acquired at similar, and relatively early, ages compared with other types of consonants. Manner has more variable acquisition with, for example, laterals and fricatives being acquired later than plosives. Based on these observations, it was predicted that manner would have more impact than place and voicing on UNWR performance because contrasts involving manner are mastered relatively late in development.

## Method

### Speech materials from Howell et al. (2017)

The UNWR task was administered to 146 children starting at five London schools in the London boroughs of Hackney and Merton (Howell et al., 2017). Ethics approval was granted by UCL's graduate school (application number 0078.004). All children in attendance in reception classes were examined providing, if they used a language other than English in the home, that language was one of those which UNWR applies to. Median age was 4.91 years and the range was from 3.98 to 6.34 years (interquartile range was 4.60–5.28 with standard error of 0.68). The gender distribution was 70 females and 76 males.

The primary languages the 146 children spoke were: English (99; 67.8%); Bengali (7 children; 4.8% of sample); Portuguese (5; 3.4%); Urdu (4; 2.7%); Lithuanian (3; 2.1%); Polish (3; 2.1%); Spanish (3; 2.1%); Turkish (3; 2.1%); Somali (2; 1.37%); and Swedish (2; 1.37%). Of the 146 participants, 51 children (i.e., 34.9% of 146) spoke a language in addition to their first language. Of those 51 children, 47 children (92.1%) spoke English as their additional language; two

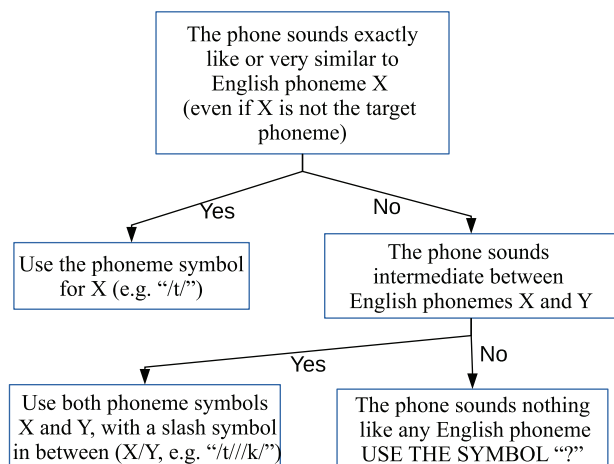
children (i.e., 3.9% of 51 children who spoke an additional language) spoke French as an additional language; one child (i.e., 1.7%) spoke Nigerian as an additional language, and one child (i.e., 1.7%) spoke Urdu as an additional language.

### Procedure for NWR testing

The target UNWR stimuli were recordings of a female phonetician using Southern Standard British English pronunciation with English stress patterns (Howell et al., 2017). These were played to children at their most comfortable volume level and the children repeated the 'made-up' words that they heard. Each test began with two-syllable long UNWR stimuli and only when all the two-syllable stimuli were presented was the syllable length increased. This same procedure was repeated for stimuli up to five syllables (maximum). There were two practices, and ten test trials per syllable length. The ten test trials at each syllable length were presented in random order and trial position was recorded for use as the control variable *trial position*. In the current study, only the two-syllable non-words were analysed because the majority of the participants (74%, 108 out of 146 participants) were not tested for non-words with three syllables or more. This was due to a stopping rule which only allowed the test to progress to the next syllable length if a child was judged to have said seven non-word test stimuli at the current syllable length correctly. Examples of the non-word stimuli included /grigre/, /plumpon/ and /blimpruk/. The participants were tested in a quiet room. Participants' were self-paced and their attempted NWRs were recorded. A Sennheiser SC 660 USB ML headset connected to a laptop was used for both the auditory presentation of the non-words and the recordings of the NWR stimuli.

### Live-scoring of NWR materials

The NWRs were scored by CS and KY for correctness live during the school testing sessions (Conti-Ramsden et al., 2001). When evaluating children's repetitions, the experimenters were instructed to focus only on the consonant errors and to ignore differences in stress, vowel length or quality (Howell et al., 2017). Each NWR was scored at the whole non-word level. All the consonants in each non-word were supposed to have been pronounced correctly by a child to be scored as correct. Conversely, a non-word should have been scored as 'incorrect', if it contained any number of consonant errors. All non-words at a given syllable length were played even when a child made an error. However, a child could progress to the next syllable length



**FIGURE 2** Decision tree for phoneme classifications. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

only when seven non-words were correct (the stopping rule referred to).

### Transcription of NWR materials

The audio responses were segmented from the recordings. All of the responses were first transcribed offline post testing by a phonetically trained research student, CS, who entered them into Praat textgrids (Boersma, 2001). The level of the transcriptions was phonemic because the aim was to maximize the reliability and consistency within and across transcribers. It is to be expected that some transcription differences are due to speech misperception and level of transcriber training. Therefore, by adopting a broader level of transcription, these influences can be moderated. APRAbet, specifically, a version used by the BEEP pronunciation dictionary (<https://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html>) was used for transcription symbols.

Only English phonemes were allowed in transcriptions. If the phone sounded exactly like, or very similar to, an English phoneme, then that phoneme was selected. If the phone sounded intermediate between two English phonemes (X and Y), then a slash symbol was used, for example, X/Y. Finally, if the phone did not sound like any English phonemes, the “?” symbol was entered. Furthermore, if the repetition contained false-starts or multiple attempts then the ‘-’ symbol was used to divide up the multiple attempts, for example, [K AE—K AE T]. These steps were represented in a decision tree to assist the transcriber (Figure 2).

To assess the reliability of the transcriptions, approximately 10% of the participants were randomly sampled from each school (18 participants were selected). The trials

RA (R) -	0.76	0.73	0.43	
KY (R) -	0.48	0.46		0.43
CS (R) -	0.84		0.46	0.73
CS (O) -		0.84	0.48	0.76
	CS (O)	CS (R)	KY (R)	RA (R)
	Transcription B			

**FIGURE 3** Pairwise Cohen’s Kappa of transcription reliability based on accuracy scores computed over the target pronunciations of the NWR stimuli.

Note: CS(O), CS’s original transcription; CS(R), CS’s retranscription; KY(R), KY’s retranscription; and KY(R), RA’s retranscription. CS(O) was used in the analyses reported.

of the 18 participants were retranscribed by the original transcriber, CS, to evaluate intra-transcriber reliability, CS(R), and by two other phonetically trained research students, KY and RA, to evaluate inter-transcriber reliability.

### Reliability analyses

Cohen’s Kappa was used as an index of interrater agreement (irr) between two raters for categorical data (Cohen, 1960). Kappas for all pairwise combinations of the transcription sets were computed using the *irr* library (Gamer et al., 2019). The pairs consisted of the original transcription by transcriber CS, and the three sets of retranscriptions by transcribers CS, KY and RA. The analyses focussed on how reliable the transcriptions were in generating NWR scores, rather than how reliable the transcriptions were between transcribers. For each set of transcriptions, their accuracies based on the target pronunciations of the NWR stimuli were computed. If a transcription matched the target pronunciation then it was considered as ‘Correct’, otherwise it was designated ‘Incorrect’. The reliability of the accuracy scores (Correct versus Incorrect) were then evaluated across transcription sets. The pairwise Cohen’s Kappas are summarized in Figure 3 and they were all highly significant. First, unsurprisingly, the intra-transcriber reliability was the highest of all comparisons (CS(O) versus CS(R):  $k = 0.84$ ) where CS(O) and CS(R) refer to CS’s original transcriptions and retranscriptions for

these materials. Of the inter-transcriber reliability scores, RA transcribed with a moderately high reliability compared with CS(O) at  $k = 0.76$ , and similar with CS(R) at  $k = 0.73$ . Transcriber KY was least reliable compared with the others with Kappa values of 0.43 and 0.48 compared with CS(O) and CS(R).

## Signal detection analyses

In all the analyses, the transcriptions were assumed to provide a good approximation to the pronunciation (see reliability statistics). To evaluate accuracy of the experimenters' live-scoring ability, offline transcriptions were examined to establish whether each trial had a consonant error or not. The transcription-based scores were then compared with the live scores in a signal detection analysis using the *neuropsychology* R library (Makowski, 2016).

## Data processing for real-time versus transcription scoring

The CS(O) transcriptions were extracted from the Praat textgrids for the 146 participants. Given that there were 12 two-syllable non-words in total, there should be 1752 trials. However, some trials were excluded from analyses: 26/1752 (1.5%) non-words did not have a live score because the participants did not attempt a repetition. Some non-words (10/1752, 0.6%) had a live score but the recordings were too poor (due to noise) or too quiet and, therefore, these non-words were not transcribed. Finally, the transcription of some non-words (74/1752, 4.2%) contained uncertain phones (namely those with a slash or a '?' symbol) or where multiple attempts were made (those with a '-' symbol). The transcription of a NWR was only extracted from textgrids if the non-words were attempted and scored live, and if the transcription did not contain uncertain phones or multiple attempts. After all these exclusions, 93.7% of the trials were retained (1642/1752).

Some of the participants were tested in two consecutive years in the Hackney school. In these cases, only the responses from the first test were used in the current analyses. Although the responses from the second test were excluded, they were nonetheless included in the automatic Levenshtein distance alignment process since the alignment quality improves with the number of items that are aligned.

The filtered dataset contained 1642 non-word attempts that had completed live scores and aligned phonemic transcriptions for the 146 participants. The distribution of participants across number of test items available are given in Table 3. Of the 146 participants, 90 participants had

TABLE 3 Distribution of the number of test items for the 146 participants.

Number of remaining test items	Number of participants
5	1
6	2
7	4
8	15
9	34
10	90

all ten items as shown in the last entry in Table 3, and 88 participants had all 12 non-word responses including the practice and test items (not shown in Table 3). The participant and item summary statistics are broken down by school in Table 4. The practice items were filtered out since the children's attempts were likely to be less robust because they were still becoming familiar with the task. 1371 non-word attempts of the test items remained and were subjected to analyses.

## Details of models

Generalized linear mixed effect logistic regression was employed using the function *glmer* from the R library *lme4*. Each model aimed to predict whether live score responses were scored correct or incorrect using a range of control variables and different explanatory variables. The three types of LMM involved different fixed effect explanatory variables but the same set of idiosyncratic random control variates (Participant, School and Target non-word). All models also included the following fixed effects as a check on whether they affected model fit: (1) Levenshtein distances *for vowels*; (2) trial position within a session (to control for fatigue and practice effects); (3) number of consonants in the target utterance; and (4) bilingualism.

For each model, correct and incorrect live scores were coded 0 and 1, respectively. Thus, a positive regression coefficient means that the likelihood of an 'Incorrect' response increases. All variables were  $z$ -transformed to allow a comparison of their relative effect sizes based on the size of their coefficients. The explanatory variables were as follows:

Model 1: Consonant Levenshtein distance.

Model 2: Consonant substitution, consonant deletion, consonant insertion.

Model 3: Consonant deletion, consonant insertion, substitution place errors, substitution manner errors and substitution voicing errors.

TABLE 4 Summary of schools from which the 146 participants attended and summary details of transcription data.

School	Number of participants	Number of participants with all test and practice items	Number of participants with all test items	Number of non-word responses (test and practice)	Number of non-word responses (test only)
H1	63	28	30	686	575
M1	31	23	23	356	298
M2	23	14	14	259	215
M3	16	12	12	187	155
M4	13	11	11	154	128

Note: H and M, Hackney and Merton boroughs and the number identifies individual schools in the boroughs.

Models 1a, 2a and 3a are the non-normalized versions and models 1b, 2b and 3b are the normalized versions. Model one served as the baseline. For model one, consonant Levenshtein distance refers to a composite score for all consonants whereas in subsequent models Levenshtein distances were computed for different types of consonant errors that occurred on materials with different phonemic properties. Model two coded consonant errors as substitutions, deletions and insertions that derive from the neighbourhood density approach and are linked to phonological processes. Model three included deletions and insertions as in model two but categorized consonant substitution errors according to place, manner and voicing confusions. Further details about the variables follow.

### Variables investigated. I. The number of consonant errors (model 1)

The number of mismatches, irrespective of whether the mismatch was a substitution, deletion or insertion was the simplest score formulated. Given the focus on consonant errors here, the Levenshtein distance of the aligned consonant segments was computed for all models. Two versions of this variable were calculated: non-normalized consonant Levenshtein distance and normalized consonant Levenshtein distance (Bailey & Hahn, 2001; Schepens et al., 2012; Wieling et al., 2014). The normalized distance was computed by dividing the non-normalized distance by the length of the alignment (Heeringa, 2004: 130–132).

### Variables investigated. II. The number of consonant substitution, deletions and insertions (model 2)

The mismatches between the actual and target pronunciations were formulated as three separate variables, namely the number of target consonant substituted (*consonant substitutions*), the number of target consonants missed (*consonant deletions*) and the number of consonants produced without a match with a target consonant (*consonant insertions*). Both normalized and non-normalized versions of these three variables were computed where normalization involved dividing the number of consonant deletions/substitutions/insertions by the length of the alignment that involved a consonant in the target or the actual pronunciation.

### Variables investigated. III. The number of place/manner/voicing feature errors (model 3)

Substitution errors differ in type and it is possible that these should not be weighted equally in overall scores (as in model 2). To examine influence of types of substitution error, they were classified into three types: place of articulation, manner of articulation and voicing. This resulted in three feature level variables, *viz* the number of place errors, the number of manner errors, the number of voicing errors. As with the other variables, both normalized and non-normalized versions of these variables were computed. The normalized variables were computed by dividing the non-normalized variables by the number of target consonants.

### Control variables investigated for all models

- I. **The number of vowel errors:** Whilst, as mentioned, the experimenters were instructed to ignore vowel errors, the number of vowel errors could still influence the live score in a number of ways. The number of vowel errors scores is the Levenshtein distance of the aligned vowel segments, irrespective of whether the error is a substitution, deletion or an insertion. Again, two versions of this variable were computed: non-normalized and normalized vowel Levenshtein distances for use in corresponding model subtypes.
- II. **The number of target consonants:** The more consonants a target non-word contains, the more likely it is that a consonant would be mispronounced. This should also increase the probability of an 'incorrect' live score.
- III. **Trial position:** The position of each non-word trial in a test session could capture potential effects of participant fatigue or practice. A positive relationship between the probability of an 'incorrect' live score and the trial position would suggest a fatigue effect. A negative relationship would suggest a practice effect.
- IV. **Bilingualism:** Children were designated as monolingual or bilingual at time of the test by consultation with the schools. Monolingual did not necessarily mean monolingual English. Children were designated monolingual if no additional language was indicated and bilingual if one or more additional language/s was indicated. This variable was contrast-coded using sum coding with the values,  $-0.5$  (reference level) and  $0.5$ , with monolingual being the reference level. Sum coding compares the mean of the dependent variable of a specific level to the grand mean of the bilingual variable.

### Model evaluation

Variance inflation factor (VIF) was computed for the variables in each model to assess effects of multicollinearity. In each model, all variables have  $VIF < 10$ , therefore they posed no issues of multicollinearity.

Model comparisons employed three metrics of model quality: Marginal  $R^2$ , Akaike information criterion (AIC) and Bayesian information criterion (BIC). Marginal  $R^2$  reflects the percentage of variance explained by the fixed effect components of a model. AIC and BIC are estimates of prediction error and BIC penalizes a complex model more than AIC. While high marginal  $R^2$  values indicate good model fit, low values of AIC and BIC indicate good model fit. AIC, BIC and Marginal  $R^2$  were computed using the *MuMIn* R library (Bartoń, 2016).

The statistical significance of the individual predictors in all the models was evaluated by bootstrapping carried out using the *bootmer* function in the *lme4* library. A total of 1000 bootstrap simulations were performed for each model. Bootstrapped  $p$ -values and 95% confidence intervals were computed for each predictor in each model. Any  $p$ -value below 0.05 is referred to as 'significant' and values above 0.05 are not significant.

## Results

### Signal detection analysis

Two metrics were computed based on signal detection theory using the distribution of hits, misses, false alarms and correct rejections in Table 5:  $A'$  is a non-parametric estimate of discriminability, and  $B''D$  is a non-parametric estimate of bias (see Pallier 2002 for the algorithms, and Macmillan and Douglas 2004 for detailed explanations of these metrics).  $A'$  of 0.83 (the non-parametric estimate of discriminability) and  $B''D$  of 0.92 (the non-parametric estimate of bias) were obtained. An  $A'$  near 1.0 indicates good discriminability, whilst a value of 0.5 signifies chance performance. A  $B''D$  equal to 0.0 indicates no bias, positive numbers represent conservative bias (i.e., a tendency to answer 'correct'), and negative numbers represent liberal bias (i.e., a tendency to answer 'incorrect'). The maximum absolute value of  $B''D$  is 1.0.

The signal detection analyses suggested that the discriminability concerning whether consonant error was present during live scoring was reasonably high with an  $A'$  of 0.83. However, there was a strong response bias towards a 'correct' live score response (the  $B''D$  value was 0.92). This indicated that the live scorers had a tendency to miss an error (a high miss rate) than to over-report an error (a low

**TABLE 5** Summary of how live scores correspond to transcription-based scores in terms of the number of hits, misses, false alarms and correct rejections (total number of scores = 1371); and all possible error types and the distribution of their correctness (based on the live scores).

	'Incorrect' live score	'Correct' live score
Consonant error present	451 (hit)	536 or 39.1% (miss)
Consonant error absent	18 or 1.3% (false alarm)	366 (correct rejection)

	Correct	Incorrect
Consonant and vowel error	159	190
Consonant error alone	377	261
No consonant or vowel error	314	11
Vowel error alone	52	7

**TABLE 6** Summary of model comparisons using Akaike information criterion (AIC), Bayesian information criterion (BIC) and marginal  $R^2$ .

Model	AIC	BIC	Marginal $R^2$ (fixed effect) (%)
1a	1462.865	1509.875	29.06%
1b	1464.543	1511.553	29.34%
2a	1457.538	1514.995	29.92%
2b (Best model)	1447.818	1505.275	31.23%
3a	1461.999	1529.902	30.27%
3b	1485.708	1553.61	29.24%

false alarm rate). Further details about error (whether both consonants and vowels, just consonants just vowels were in error or there was no error at all) are presented in the bottom section of Table 5.

## LMM model outcomes

Table 6 summarizes the marginal  $R^2$ , AIC and BIC metrics for the six models. Model 2b (normalized substitution, deletion, insertion model) was the best model on all three metrics as it had the highest marginal  $R^2$  and the lowest AIC and BIC values (Table 7). All the fixed factors apart from trial position and bilingualism were highly significant.

For model 2b, the consonant error variables suggested that a higher proportion of consonants substituted, deleted or inserted, increased the likelihood of an 'incorrect' live score. The relative effect sizes based on the size of their coefficients showed that consonant substitution had the strongest effect ( $\beta = 0.8941$ ), followed by consonant deletion ( $\beta = 0.5336$ ) and then consonant insertion ( $\beta = 0.5055$ ).

The sign of the significant  $\beta$  coefficients of the control variables were positive both for number of target consonants and vowel Levenshtein distance with values of 0.3710 and 0.2468, respectively. The number of consonants

control variable suggested that a higher number of target consonants increased the likelihood of an 'incorrect' live score. Similarly, the vowel error control variable suggested that a higher proportion of vowels substituted, deleted or inserted increased the likelihood of an 'incorrect' live score. The coefficients for the non-significant control variables trial position and bilingualism variable were  $-0.0326$  and  $-0.2882$ .

## General discussion

The re-analysis of Schwob et al.'s (2021) systematic review, which examined procedures used in NWR research studies, showed that a third of studies scored judgments live. Thus, the significant proportion of research studies that were based on live scoring lacked recordings for analyses to establish reliability and validity. Moreover, when recordings are absent, it is not possible to determine how the accuracy of results obtained when live scores are made compare with those obtained using more involved procedures. Consequently, it is not clear how well issues in phonological working memory are identified using live procedures.

Experiment 2 was designed to provide information about how live judgments compared with ones based on transcriptions of the responses to begin to establish the

TABLE 7 Fixed effects summary for Model 2b.

	$\beta$	SE	Z	CI (lower 95%)	CI (upper 95%)	p-value
(Intercept)	-1.0556	0.2134	-4.9494	-1.4625	-0.6225	< 0.001***
Consonant substitution (normalized)	0.8941	0.0764	11.7011	0.7393	1.0414	< 0.001***
Consonant deletion (normalized)	0.5336	0.0700	7.6209	0.3841	0.6752	< 0.001***
Consonant insertion (normalized)	0.5055	0.0686	7.3698	0.3713	0.6377	< 0.001***
The number of target consonants	0.3710	0.1319	2.8119	0.0934	0.6372	0.002**
Vowel Levenshtein distance (normalized)	0.2468	0.0697	3.5609	0.1044	0.3782	< 0.001***
Trial position	-0.0326	0.0670	-0.4866	-0.1754	0.1101	0.710 (n.s.)
Bilingual (multilingual versus monolingual (reference level))	-0.2882	0.1561	-1.8463	-0.6138	0.0455	0.080 (n.s.)

Note: Variables are presented in descending size of  $\beta$ -values.  $\beta$ , coefficient; SE, standard error; z, z-value; CI (lower 95%) and CI (upper 95%), 95% confidence intervals of the coefficient from bootstrapping; p, p-value from bootstrapping simulations.

extent to which live scores are sensitive to phonological working memory issues. It is acknowledged that speech and language therapists (SaLTs) may not have time to do transcriptional analyses routinely. However, they need to know the limits of live procedures if they use them so that they are aware what errors they are most likely to miss. By providing this information, this study indicated how their work may be improved in the future (e.g., by designing procedures that allow them to focus attention on errors they are likely to miss).

Accuracy of NWR live-scoring relative to transcriptional procedures was investigated for a particular NWR test (UNWR). This test is appropriate for groups of children with various language backgrounds (Howell et al., 2017) which was the case with the cohort tested. Choice of what measures to use for assaying errors in live scores was made based on parameters that: (1) were appropriate to characterize all types of errors and (2) linked to errors made in speech development (problems with phonological processes and difficulties with particular phonemes; Dodd, 2005; Harris, 1994). It was not possible to focus analyses on particular phonological processes and specific problem phonemes (Dodd, 2005) given current disagreements about findings and incomplete statistics. Instead, the current analyses used word-usage operators (substitution, deletion and insertion) that are involved in typical and atypical phonological processes (Table 2) and traditional descriptors that are appropriate to all phonemes (place, manner and voicing) which relate to phonemic difficulty.

Three types of models were fitted to the data to determine factors responsible for sources of discrepancy between live scores and transcriptions. The first used error scores on any consonant in a stimulus and was the baseline for comparison with the other two model types. Model type two fitted data to assess the impact of usage parameters alone (these link to phonological processes). Model type

three fitted usage and phonemic factors (the latter links to difficulties on particular sounds). All three types of models had normalized and non-normalized variants.

Model 2b that used normalized substitution, deletion and insertion metrics that are employed in the neighbourhood density field provided the best model for accounting for discrepancies between live and transcription-based scores. The fits with this model were better than when only number of consonants (model 1) or when substitutions were broken down further into place/manner/voicing dimensions (model 3). This suggests that issues in assessing phonological processes rather than problems with particular sounds need attention when live judgments are made.

Models 1a and 1b provided benchmarks which just used number of consonants as measures of the discrepancy between children's performance and the target transcriptions and their use as benchmarks was supported in the analyses: Each of these models had higher prediction errors (AIC and BIC) and accounted for less variance than their corresponding type 2 models. Comparison of corresponding type 1 and 3 models showed that prediction errors for type 3 models were superior. For the non-normalized type 3 model, variance accounted for (marginal  $R^2$ ) was comparable to type 2 models.

In model 3, substitution errors were divided into place, manner and voicing confusions. This did not improve the model significantly over the simple consonant substitution error variable used in type 2 models. The lack of place/manner/voicing effects suggests that the live scorers were not sensitive to errors at this higher level of detail. Its absence might arise due to an instructional effect. Thus, the live scorers were told to identify whether the non-word was produced correctly or incorrectly with respect to the consonants' identity and position alone, thereby ignoring any potential vowel differences. This could lead live scorers to focus on types of consonant error (substitution,





deletion or insertion) at the expense of types of substitution (in terms of place/manner/voicing). If instructions given to scorers had specifically asked them to pay attention to place, manner, and voicing of consonant substitutions, model 3 might have been appropriate.

Two of the control variables significantly affected results for model 2b (Table 7). First, the number of consonants in a word was a significant variable, although consonant errors normalized by the number of target consonants was included as a predictor. This may suggest that a higher number of consonants in a non-word increases the likelihood of live scorers scoring the non-word as 'incorrect', regardless of whether consonant errors were present or not. This might arise because judges had to hold a higher number of consonants in working memory whilst scoring. Alternatively, it is possible that scorers held prior expectations such that non-words with a higher number of consonants would be more likely to be erroneous, hence biasing their scoring accuracy.

Second, the number of vowel errors was not expected to have an effect but it did in all models. Examining the transcription data further, there were 59 responses that had no consonant errors but where there was a vowel error. Of these, 52 (88%) were given 'Correct', and 7 'Incorrect', live responses (bottom section of Table 5). From this it appears that the live scorers were good at ignoring vowel errors when only the vowel was in error. Alternatively, the number of vowel errors could influence the live score in a number of ways. First, live scorers may not be able to ignore vowel differences at all times whilst making scores. Second, vowel differences may lead to 'perceived' consonant differences, where for example an incorrectly produced vowel adjacent to a correctly produced consonant may cause the consonant to be perceived differently and therefore scored as incorrect. Perceptual repair due to top-down biases may be a possible mechanism behind this account. For example, a vowel error might create a phonological environment (such as CVC) that is impossible or highly unlikely in a given language in terms of its 'phonetic naturalness' (Hayes & White, 2013), therefore causing a perceptual repair which changes the environment (e.g., consonants) to fit better with the produced vowel. This relates to the idea of substantive bias in phonotactics, whereby it is assumed that even in artificial words, there is a preference for perceptually similar sounds (White, 2014; White, 2017). Third, vowel differences may lead to adjacent consonants being produced slightly differently because participants co-articulated the consonants with the vowels. The production differences of the consonants might not be different enough to be perceived as a different phoneme by the offline transcribers, but the differences could be perceived as a different phoneme during live judgements. Further work is needed to determine which, if any, of these possibilities applies. Finally in connection with the control

variables, although not significant, including bilingualism compared with identical models that did not include this factor, improved model 2b's fit slightly.

## Limitations and future work

The re-analysis of the systematic review in Experiment 1 showed that recommendations are required about what procedural details should be reported in studies. This is not to suggest one format would be appropriate. Whilst live scoring could be allowed when transcriptions are made, the level of transcription that was attempted (orthographic or phonemic etc.) should be specified. Also, different performance metrics could be allowed (marginal  $R^2$ , AIC and BIC were reported here) including those preferred by other authors. Transcription-based research studies differ from work where judgments are made live in that the former is based on audio records and typically analyses parts of words whereas the latter usually has audio *and* visual information about the speaker and a whole-item judgment is made. All these factors should be recorded when procedures are described. Further documentation is needed when transcriptional analyses were made since Experiment 1 showed that there was considerable variability in what was done (there, the only detail that was documented was whether whole-item or part-item analyses were conducted).

Experiment 2 raised further issues about recommendations when reporting transcription-based analyses. Were consonants, consonants and vowels or holistic judgments made and were each of these scored and assessed for reliability and validity? This would not only improve standard of reporting in research studies, but also provide clear benchmarks against which to adjudge clinical assessments using live scoring.

There were limitations in the procedures adopted here. For example, in Experiment 2, even though a detailed transcription guide was available and a phonemic-level of transcription alone was required, the reliability analyses showed that there was, nonetheless, considerable variability between transcribers. A second limitation is that only one NWR format (UNWR) was examined. Future studies need to examine other NWR formats such as those created using the CLT (Chiat, 2015) and, for all NWR formats, assessments need to extend to syllable lengths longer than two syllables. General phonemic descriptors were used here (e.g., place, manner and voicing). An alternative would be to define phonemes that are difficult to produce based on particular manner classes (e.g., fricatives and laterals).

There are also issues about the measures used in Experiment 2. Whilst judgements of substitutions, deletions and insertions provided good measures for assessing difference

in scores between those made live versus those based on transcriptions, the case that the usage parameters relate to typical, delayed and atypical phonological processes in development (Table 2) needs further investigation. Whilst a possible advantage is that the usage measures characterize all phonological processes objectively, they may lack sensitivity when individual children have difficulty with particular processes which may be especially important for the atypical processes associated with language disorder (Dodd, 2005). Future work should address this. Irrespective of whether clinical conditions involve phonological delay or disorder (Howell, 2010), reliable and objective procedures need to be available to establish departures from patterns of typical development (slowed or atypical).

The current procedures could be used as a taxonomy for characterizing various speech and language conditions in future work. Stuttering, for instance, might (hypothetically) be considered to be a speech motor output condition with intact language functioning. If that position is taken, then deletions would be a way that a child who stutters might tackle this issue whereas insertions and substitutions would be less favoured responses (a different pattern to that described for the current sample where substitutions dominated). Also, particular phone types involving place, manner and voicing may be more susceptible to substitution for people who stutter. For instance, voicing is a well-documented feature that people who stutter find particularly problematic (Howell & Vause, 1986; Howell & Williams, 1988, 1992).

There are limitations in the sample used. To date, assessments have only been made on non-clinical samples of school-age children. Experiment 2 permitted heterogeneity of language background by employing the UNWR test. However, children from clinical groups and older participants also need assessing on UNWR and other NWR tests. This should include assessment of whether the scoring procedures used here are appropriate for participants from different demographic groups.

So far in this report, UNWR has been considered as a test of language development. However, NWR tests are used in assessments of cognitive as well as language issues and language may be less affected in the former. An example is ADHD where phonological working memory issues have been reported in NWR tasks (Kasper et al., 2012) but where, a priori, the visual encoding or executive function components of working memory may be more affected than language processes (Raiker et al., 2017). If the current scoring metrics do not work for ADHD participants, scoring metrics for these other components would need developing.

Currently, a major challenge is how to provide the detail available in the manual transcriptions (these are prohibitively expensive to conduct routinely in clinical

assessments). In future work, there are a range of machine learning technologies that are being used for various applications in speech therapy (Barrett et al., 2022) that could be relatively easily adapted to provide transcriptions of NWR data. Transcriptions such as those available from this study could be used to train automatic speech recognizers (ASRs) to give a procedure for the phonemic transcripts of children's speech. Training ASRs needs considerable amounts of data which, in itself, incurs significant cost. Nevertheless, this issue should be more tractable than, for instance, ASR of stuttered speech even though this field has seen considerable developments in the last 25+ years (Barrett et al., 2022). The reasons why automatic UNWR scoring should be comparatively easy are: (1) there are a small finite number of known targets being attempted; and (2) recordings are usually made for validation of live scores with similar live scores made from recordings. Such recordings could be made available in online repositories for preparation of training materials as done, for example, with the UCLASS database of stuttered speech (Howell et al., 2009).

## Conclusions

The main study (Experiment 2) showed that improved NWR scoring procedures are required to provide SaLTs with better indications about children's language development (typical and atypical). The procedures currently used miss, in order of increasing likelihood of live score errors, substitutions, deletions and insertions. Hence, procedures are required that provide the information currently only available when conducting transcriptions manually which is not feasible in clinical work. The possibility of training automatic speech recognizers to provide this level of detail was raised to address how to make procedures feasible to conduct and to provide standard analyses.

## ACKNOWLEDGEMENTS


Open access funding enabled and organized by Projekt DEAL.

## DATA AVAILABILITY STATEMENT

The data and code for the analyses in Experiment 2 can be accessed at the Open Science Framework repository <https://www.doi.org/10.17605/OSF.IO/75AE3>.

## ORCID

Peter Howell  <https://orcid.org/0000-0001-5361-5031>

Roa'a Alsulaiman  <https://orcid.org/0000-0002-0828-5181>

John Harris  <https://orcid.org/0000-0002-1675-6860>

Kevin Tang  <https://orcid.org/0000-0001-7382-9344>

## REFERENCES

- Archibald, L.M. (2008) The promise of nonword repetition as a clinical tool. *Canadian Journal of Speech–Language Pathology and Audiology*, 32(1), 21–28.
- Barrett, L., Hu, J. & Howell, P. (2022) Systematic review of machine learning approaches for detecting developmental stuttering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 1160–1172.
- Bailey, T.M. & Hahn, U. (2001) Determinants of wordlikeness: phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44(4), 568–591.
- Bartoń, K. (2016) Multi-Model Inference (MuMIn). R package version 1.15.6.
- Boersma, P. (2001) Praat, a system for doing phonetics by computer. *Glott International*, 5:9/10, 341–345.
- Chiat, S. (2015) Non-word repetition. In: Armon-Lotem, S., de Jong, J., & Meir, N. (Eds.) *Methods for assessing multilingual children: disentangling bilingualism from language impairment*. Bristol, United Kingdom: Multilingual Matters, pp. 125–150.
- Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Conti-Ramsden, G., Botting, N. & Faragher, B. (2001) Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry*, 42 (6), 741–748. <https://doi.org/10.1017/S0021963001007600>
- Crowe, K. & McLeod, S. (2020) Children's English consonant acquisition in the United States: a review. *American Journal of Speech–Language Pathology*, 29(4), 2155–2169.
- Daland, R. (2015) Long words in maximum entropy phonotactic grammars. *Phonology*, 32(3), 353–383.
- Dockrell, J. & Howell, P. (2015) Identifying the challenges and opportunities to meet the needs of children with Speech Language and Communication Difficulties. *British Journal of Special Education*, 42, 411–428.
- Dodd, B. (Ed), (2005) *Differential diagnosis and treatment of children with speech disorder*. 2nd edition, Chichester: Whurr.
- Dollaghan, C. & Campbell, T.F. (1998) Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, 41, 1136–1146.
- Gamer, M., Lemon, J., Fellows, I. & Singh, P. (2019) irr: Various coefficients of interrater reliability and agreement. R package version 0.84.1. <https://CRAN.R-project.org/package=irr>
- Gathercole, S.E., Willis, C.S., Baddeley, A.D. & Emslie, H. (1994) The Children's Test of Nonword Repetition: a test of phonological working memory. *Memory (Hove, England)*, 2(2), 103–127. <https://doi.org/10.1080/09658219408258940>
- Hayes, B. & White, J. (2013) Phonological naturalness and phonotactic learning. *Linguistic Inquiry*, 44(1), 45–75.
- Harris (1994) *English sound structure*. London, United Kingdom: Blackwell.
- Heeringa, W. (2004) *Measuring dialect pronunciation differences using Levenshtein distance*. Doctoral dissertation. Groningen: University of Groningen.
- Heeringa, W., Gooskens, C., Nerbonne, J. & Kleiweg, P. (2006) Evaluation of string distance algorithms for dialectology. In: Nerbonne, J. & Hinrichs, E. (Eds.) *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), pp. 52–62.
- Howell, P. (2010) *Recovery from Stuttering*. New York: Psychology Press.
- Howell, P., Davis, S. & Bartrip, J. (2009) The UCLASS archive of stuttered speech. *Journal of Speech, Language and Hearing Research*, 52, 556–569. [https://doi.org/10.1044/1092-4388\(2008/07-0238](https://doi.org/10.1044/1092-4388(2008/07-0238)
- Howell, P., Tang, K., Tuomainen, O., Chan, K., Beltran, K., Mirawdeli, A. & Harris, J. (2017) Identification of fluency and word-finding difficulty in samples of children with diverse language backgrounds. *International Journal of Language & Communication Disorders*, 52, 595–611.
- Howell, P. & Vause, L. (1986) Acoustic analysis and perception of vowels in stuttered speech. *Journal of the Acoustical Society of America*, 79, 1571–1579.
- Howell, P. & Williams, M. (1988) The contribution of the excitatory source to the perception of neutral vowels in stuttered speech. *Journal of the Acoustical Society of America*, 84, 8089.
- Howell, P. & Williams, M. (1992) Acoustic analysis and perception of vowels in children's and teenagers' stuttered speech. *Journal of the Acoustical Society of America*, 91, 1697–1706.
- Kasper, L.J., Alderson, R.M. & Hudec, K.L. (2012) Moderators of working memory deficits in children with attention-deficit/hyperactivity disorder (ADHD): a meta-analytic review. *Clinical Psychology Review*, 32, 605–617.
- Macmillan, N.A. & Creelman, D. (2004) *Detection theory: A user's guide*. Psychology Press.
- Makowski, D. (2016) Package 'neuropsychology': An R Toolbox for Psychologists, Neuropsychologists and Neuroscientists. Available from: <https://github.com/neuropsychology/neuropsychology.R>
- Masoura, E.V. & Gathercole, S.E. (1999) Phonological short-term memory and foreign language learning. *International Journal of Psychology*, 34(5–6), 383–388.
- Nerbonne, J., Heeringa, W. & Kleiweg, P. (1999). 'Edit distance and dialect proximity.' *Time warps, string edits and macromolecules: the theory and practice of sequence comparison*. In: Sankoff, D. & Kruskal, J. Stanford, CA: CSLI Publications.
- Oh, Y.M., Pellegrino, F., Coupé, C. & Marsico, E. (2013) Cross-language comparison of functional load for vowels, consonants, and tones. *Interspeech*, 3032–3036).
- Oh, Y.M., Coupé, C., Marsico, E. & Pellegrino, F. (2015) Bridging phonological system and lexicon: insights from a corpus study of functional load. *Journal of Phonetics*, 53, 153–176.
- Pallier, C. (2002) Computing discriminability and bias with the R software. Available from: <https://www.pallier.org/pdfs/aprime.pdf>
- Raiker, J.S., Friedman, L.M., Orban, S.A., Kofler, M.J., Sarver, D.E. & Rapport, M.D. (2017) Phonological working memory deficits in ADHD revisited: the role of lower level information-processing deficits in impaired working memory performance. *Journal of Attention Disorders*, 23(6), 570–583. <https://doi.org/10.1177/1087054716686182>
- Schwob, S., Eddé, L., Jacquin, L., Leboulanger, M., Picard, M., Oliveira, P.R. & Skoruppa, K. (2021) Using nonword repetition to identify developmental language disorder in monolingual and bilingual children: a systematic review and meta-analysis. *Journal of Speech, Language and Hearing Research*, 64, 3578–3593.
- Schepens, J., Dijkstra, T. & Grootjen, F. (2012) Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, 15(1), 157–166.



- Tang, K. (2015) 'Naturalistic Speech Misperception'. PhD thesis. University College London.
- White, J. (2014) Evidence for a learning bias against saltatory phonological alternations. *Cognition*, 130(1), 96–115.
- White, J. (2017) Accounting for the learnability of saltation in phonological theory: a maximum entropy model with a P-map bias. *Language*, 93(1), 1–36.
- Whiteside, S.P. (1996) Temporal-based acoustic-phonetic patterns in read speech: some evidence for speaker sex differences. *Journal of the International Phonetic Association*, 26(1), 23–40.
- Wieling, M., Bloem, J., Mignella, K., Timmermeister, M. & Nerbonne, J. (2014) Measuring foreign accent strength in English: validating Levenshtein distance as a measure. *Language Dynamics and Change*, 4(2), 253–269.
- Wieling, M., Prokić, J. & Nerbonne, J. (2009) Evaluating the pairwise string alignment of pronunciations. In: Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education. Association for Computational Linguistics, 26–34.
- Windsor, J., Kohnert, K., Lobitz, K. & Pham, G. (2010) Cross-language nonword repetition by bilingual and monolingual children. *American Journal of Speech–Language Pathology*, 19(4), 298–310.

**How to cite this article:** Howell, P., Sorger, C., Alsulaiman, R., Yoshikawa, K., Harris, J. & Tang, K. (2023) Factors affecting judgment accuracy when scoring children's responses to non-word repetition stimuli in real time. *International Journal of Language & Communication Disorders*, 1–20. <https://doi.org/10.1111/1460-6984.12954>