CORTEX 205 (2026) 241–254

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Methods and Assumptions

What counts as an error in nonword reading? A methods study of acquired phonological dyslexia

Jeremy J. Tree [a,b,*], *Veronika Zakharchenko* [a], *Alex L. Jones* [a] *and Benjamin A. Jones* [a]

[a] *Department of Psychology, Faculty of Medicine, Health and Life Sciences, Swansea University, Swansea, UK*
[b] *Department of Psychology, Brunel University of London, Kingston Lane, Uxbridge, Middlesex, UB8 3PH, UK*

ARTICLE INFO

Article history:
Received 13 April 2026
Revised 16 July 2026
Accepted 2 September 2026
Action editor Holly Robson
Published online 24 September 2026

Keywords:
Acquired dyslexia nonword reading

ABSTRACT

Acquired phonological dyslexia (PD) is typically characterised by impaired nonword reading, but defining what constitutes a nonword reading "error" is challenging due to variability in how even healthy readers pronounce nonwords. This study examined different methods for classifying such errors using data from younger adults ($n = 41$), older adults ($n = 25$), and a case series of individuals with suspected acquired PD ($n = 6$). Three scoring approaches were compared: (1) a mode-response accuracy method, which matches responses to the most common pronunciation; (2) an average response-overlap method, which assesses similarity to the full distribution of control responses; and (3) a singularly distinct response count, which identifies responses not produced by any other participant. Results showed that older adults exhibited greater variability than younger adults across all measures, suggesting that age significantly affects nonword reading performance. This highlights an important methodological issue: using younger normative data to assess older individuals may overestimate impairment. In the patient group, all three scoring methods broadly agreed in ranking severity across cases. However, the singularly distinct response metric provided clearer separation between patients and controls and captured a wider range of impairment severity. It also revealed that unique responses are relatively common even among typical readers, an underexplored finding. Overall, the study supports the importance of age-appropriate norms and more nuanced, response-level scoring methods in assessing acquired PD, with singularly distinct responses offering a potentially sensitive marker of disruption.

© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

* *Corresponding author.*
E-mail address: Jeremy.Tree@brunel.ac.uk (J.J. Tree).
https://doi.org/10.1016/j.cortex.2026.09.012
0010-9452/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Learning to read is a fundamental aspect of human language development and a vital skill for day-to-day life in modern society. Essentially, learning to read involves understanding how to map strings of letters (graphemes) of a newly learnt word to their corresponding speech sounds (phonemes), a process that can be somewhat replicated and observed by presenting people with entirely new letter strings ('nonwords') to read. In cognitive models of human reading, the process by which a nonword is read aloud draws upon previously learnt knowledge about how letters and sounds correspond (known as grapheme to phoneme correspondence – Coltheart et al., 2001). For example, when shown the nonword SHOIL, you might draw on elements of the words SH-E and OI-L – stitching together the pronunciation "SH-OI-L". Put simply, the process of learning to read involves acquiring the general 'rules' of the units of spelling to sound correspondence, and the process by which we respond to new words (nonwords) is a process of drawing on this increasingly developed set of rules.

Illustratively, during the period in which children develop the ability to read accurately, errors can occur within this rule-based system due to overgeneralisation. For example, once one has learnt to pronounce the English words 'mint', 'hint' and 'glint', a rule will form about a consistent way to pronounce the string '-int' – which is why a child may read the irregular word 'pint' to rhyme with 'hint' (an example of misattributing a learnt rule). Importantly, patients with brain injury can make the same kinds of errors with irregular words (surface dyslexia – e.g., Ktori et al., 2016, Playfoot, Coltheart, Saunders, & Tree, 2010) whilst others struggle to read nonwords at all (phonological dyslexia – e.g., Friedman, 1995; Tree & Playfoot, 2015) such that they often produce real word ('lexicalisation') responses (e.g., Nool = "Know", Hoz - "Whose"). This is *despite* being instructed that the items presented are not real words! In both cases this is interpreted as reflecting the fact that the brain injury has disrupted the previously 'normal' reading system in different ways.

Acquired phonological dyslexia (PD - Beauvois & Derouesne, 1979; Sartori et al., 1984) is the focus of the current work, particularly how best to classify observed performance reading nonwords as 'impaired' or otherwise. As has already been mentioned, some patients make 'lexicalisation' (real word) errors, and the fact that patients cannot help making responses despite knowing the items are nonwords would clearly constitute an unambiguous reading 'error'. After all, cognitively intact normal readers would not be expected to make such errors. However, not all patients' responses made to nonwords are necessarily quite so cut and dry in classification. That is, a perennial issue in this context remains regarding how best to decide a particular reading response to a nonword is indeed an 'error', when there is no a priori accepted consensus on how to pronounce that nonword. Suppose one assumes a nonword may have more than one observed response in the general 'healthy' reading population (akin to many words). How does one best interpret observations with PD cases? These issues are important, given that in many cases, published work on PD has informed theories of reading in critically keyways. For example, Tree and Kay (2006) examined the relationship between 'general' phonological impairment and poor nonword reading; while Tree (2008) reviewed the cases of PD concerning co-occurring impairments of NW repetition and/or function word reading. Moreover, the principle of a 'non-lexical' (nonword) reading route is captured in the classic cognitive model of reading outlined in the Dual-route model of reading (DRC - Coltheart, et al., 2001) with the assumption that damage to this route results in the kinds of performance observed in PD (c.f., Nickels, et al., 2008 for examples of computational implemented 'lesion' work). Consequently, it is apparent that if we wish to use the evidence provided by the

phonological dyslexic population to inform theories like these, we must resolve the problem of how to classify responses as 'errors', and this will be the focus of the current work.

As a first step to understanding what constitutes an 'error' in nonword reading, it is critical to consult normative data from healthy adult readers (Tree, 2008). Generally, the most typical approach that has been utilised in this respect is to consider the modal response to a nonword - put simply, if we ask 30 people to name a particular nonword, what is the most frequently observed response? However, as has been already suggested, several studies have noted that many nonwords, when read aloud by different people, yield more than one observed pronunciation across individual readers (see also, Andrews & Scarratt, 1998; Glushko, 1979; Funnell, 1983; Masterson et al., 1995; Seidenberg et al., 1994), making the 'modal' accuracy approach more or less representative depending on the item. The fact that nonwords may generate the same response across all the samples tested, whilst others can have a high number of different provoked responses, has been dubbed variation in response agreement (RA) (Mousikou et al., 2017). Put simply, different nonwords will be more or less well represented by a modal approach for scoring accuracy, and this is most likely why those interested in early screening PD cases initially, often ensured the use of a set of NWs that they had pre-established had very high levels of response agreement (i.e., $N = 1$), as a sensible first step in test development. A pertinent example of this is the Psycholinguistic Assessments of Language Processing in Aphasia (PALPA) battery (Kay et al., 1996), which has a nonword reading test comprising 24 items, all of which have an RA approaching 1 (i.e., items for which a very high majority of participants generated a singular specific response). Under this approach in effect, the 'modal response' is also the agreed response, and so a clinician can simply score errors if the patient's responses differ. In the case of the PALPA NW test, an inspection of the manual finds that they reported testing a normative sample of 32 subjects and mean 'accuracy' = 22.88 (SD = .72). Thus, even with this group of cognitively intact participants, it was the case that on average at least one response for each participant was a non-modal response. We will return to this issue later. In any case, clinicians might typically then calculate a two-standard deviation 'cut-off' for classifying impaired performance - one can thus calculate that typically 4 or more 'error' responses would meet criteria. To foreshadow the current work, we (a) have used this test to initially screen our case series of PD candidates and (b) will

investigate the utility of a modal response approach in the classification of our PD cases, whilst seeking to compare it to two alternative approaches for classifying NW impairment severity.

A possible limitation of the PALPA's 'modal' approach is that, by selecting only high-agreement items, it narrows the stimulus set. This may then hide some of the natural variability present across the wider set of possible nonwords. For example, Pritchard et al. (2012) compiled a corpus of 412 monosyllabic nonwords read by 45 adult Australian readers, aiming to compare real human responses to those generated by computational cognitive models of reading (DRC *vs* CDP+). Previous comparisons with computational models focused on whether the model's output matched the most frequent (modal) human response. However, Pritchard and colleagues suggested an alternative approach, which sought to determine the degree of overlap between a model's set of responses to the observed responses of every reader in their sample of 45 participants. Put simply, for a given individual, we do a one-to-one comparison of the responses for each of the 45 participants for their 412 nonwords and simply count the number that are identical to generate a proportion overlap. In this case, they found that the DRC model generally had a much higher average overlap than the CDP model, and thus the former was favored. Thus, in the example of this work, the primary goal was to evaluate how closely these models mirrored human behavior, but in our case their method inspires a similar approach for studying PD (see also Coltheart & Ulicheva, 2018). Importantly, by comparing each PD's responses to every other individual healthy adult reader in a key normative sample, one can derive an average response overlap. This distributional measure allows us to potentially identify divergent individual performance, by determining if the overlap levels are much lower than that typically observed in the normative sample. We will therefore adopt this more inclusive, distributional approach in our analyses of PD nonword reading and contrast this with the more 'classical' modal response scoring method.

However, implicit in both these approaches is the assumption that responses generated will likely reflect some 'subset' of possible responses (e.g., perhaps the nonword may be associated with four distinct or 'unique' responses), with the likely proportion of these 'unique' responses varying as a function of this probability-driven GPC rules system. What is not expected is that individual healthy participants might regularly generate responses that are never observed in the healthy reader database; we dub these singularly distinct responses given they are unique to only a single participant. Although it is implicit that healthy readers shouldn't behave this way, PD cases, on the other hand, if one assumes the GPC system has been severely 'disrupted', would be expected to make such responses. Put simply, brain injury (and impairment) should result in some kind of qualitative shift in response pattern that would result in singularly distinct nonword responses; in the same way as the observation of 'real word' responses (mentioned earlier) are often observed for PD cases (and nearly never for healthy readers). Thus, an important prediction concerning PD 'errors' in nonword reading is that they should likely generate lots of singularly distinct responses. However, this raises the question, do healthy adult readers ever act this way, or is it always the case that for any set of observed nonwords responses for a given individual some proportion of singular distinct responses emerge?

Surprisingly, this key distinction has been largely overlooked by the field. Our inspection of the database produced by Pritchard and colleagues indicates that such singularly distinct responses are indeed observed in healthy adult readers - in fact, 1823 of the observed responses for the 412 Nonwords tested were such! Across participants, the proportion of such responses varied from 1% to 34.5% (mean: 9.83% standard deviation: 7.16%), which suggests individuals can considerably vary in this respect. But the critical point is that all participants generated some number of provoked responses that only they generated. This actually has quite profound consequences for how we might interpret what 'abnormal' performance in PD might constitute, since it captures a qualitatively different response characteristic often overlooked. To reiterate, if any given individual reader of a corpus of nonwords can generate responses that do not match any response in a normative sample, how can one determine what 'abnormal' looks like? Put simply, one cannot merely define PD performance as reflecting the presence of some 'odd' responses, where 'odd' constitutes something nobody else (in the tested population) generates! As a consequence, we will adopt a third approach of analysing our PD case nonword responses, which will focus on the comparative proportion of singularly distinct responses, to ask whether individual PD cases indeed show disproportionately large numbers of such 'errors'?

Overall, the current project thus sought to contrast three different normative sample approaches to analysing individual patient nonword reading performance by: (a) classifying observed responses as 'correct' if they match the 'mode' response for a control sample of healthy readers (an approach typically used and discussed earlier), (b) by comparing responses of PD cases to each individual healthy adult reader to determine the degree of response overlap and calculate an average similarity proportion for individual patients (inspired by the work of Pritchard and colleagues); this attempts to more realistically capture the full range of responses that might occur in normal readers (rather than the single most dominant response), (c) counting the number of observed singularly distinct responses individual PD cases make to determine if this number is much greater than is typically observed. Suppose one assumes that acquired phonological dyslexia leads to a disruption of the cognitive reading system that entails a qualitative shift in responses. In that case, we can assume that the relative increase in such singularly distinct responses would thus reflect some degree of severity of damage to this system. Generally, impairments in neuropsychology are measured by the number of errors made; in this case, we are effectively comparing different means of determining what 'errorful' performance implies in PD. Put simply, do all three of these approaches converge in judging impairment severity in our case series or might they differ?

So far, our discussion of analysing PD nonword reading has discussed comparing different approaches that focus on responses generated by individuals, However, it is important to stress that distributions of observed responses for nonword

reading also vary by individual items. Mousikou et al. (2017) published a huge database of 916 nonword items tested across 41 healthy control participants (more than 37,500 responses!). They sought to explore the degree of observed response variance linked to items to determine what variables may predict variability, which they called response agreement (RA), which we discussed earlier, and strikingly RA values differ considerable ranging from RA = 1 to as high as 22. Therefore, nonwords items themselves may likely be well suited to be used as test materials in the identification of PD reading. To foreshadow the current work, we selected a subset of 201 nonwords from the database provided by Mousikou and colleagues that captures this range in RA (i.e., from 1 to 22). Finally, given that PD predominantly affects older adults, we address a significant gap in the literature by collecting normative nonword reading data from an older adult sample. Previous studies, such as those by Pritchard et al. (2012) and Mousikou et al. (2017), have focused exclusively on younger populations (i.e, undergraduates), leaving unanswered questions about how age influences nonword reading variability. By comparing response distributions across age groups, we aim to determine the extent to which age-related changes affect the suitability of normative data for evaluating PD, and the degree to which normative 'cut-offs' need to reflect this fact. Together, these analyses seek to advance our understanding of nonword reading variability, improve methods for classifying errors in PD, and ultimately enhance the utility of PD cases for informing cognitive theories of reading.

2. Methods

2.1. Materials

2.1.1. Young adult reader reference sample

Our stimulus set was drawn from the work of Mousikou et al. (2017) who transcribed the responses of 41 participants (UG students aged 19–27, SD = 1.6) reading 915 nonword items, resulting in a database of more than thirty-seven thousand participant responses - and our inspection of this database determined that different NWs have fewer/greater numbers of associated different 'unique' responses (i.e., NW X is associated with A, B or C response), that is they varied in what Mousikou and colleagues dub response agreement. As mentioned earlier, it would be unfeasible to use the full 915 NW set with our PD patients, so we initially pre-selected a smaller subset of 201 items that represented the varied degree to which this range was observed (i.e., initial sampling ensured selection of NWs that encompassed the full continuum on this variable). Thus, from the initial younger adult control sample provided by Mousikou and colleagues, we have 41 participants and 201 NWs, constituting 8241 (41 × 201) trial responses – below we discuss how successful we have been in ensuring this sub-samples largely reflects similar properties to the larger full sample of 915 NWs.

Across this sub-sample database, we observed a total of 1397 different 'unique' responses. Appendix 1 provides the list of nonwords (see OSF: https://doi.org/10.17605/OSF.IO/XMZ49), with each of the responses associated with that nonword and the range of the counts of these stimulus-specific responses (e.g., 10/201 provoked only one consistent response, whilst in contrast, two items provoked a high count of different responses: SHERUCH = 21, ELUCH = 22). Importantly, our sub-sample varies not only on response agreement, as operationalised by the count of 'unique' responses observed for a given item, but also the proportions of these responses illustrate quite different patterns (see Appendix 1). For example, in some cases, a nonword may be associated with three responses where one is very strongly dominant, whilst in other cases the distribution of responses is much more balanced (i.e., no single response emerges as a 'winner'). In Appendix 1, we have categorised the NWs into two types, on the one hand, 'dominant' response NWs are characterised by one response that accounts for at least 70% of all observed responses, whilst anything less is characterised as 'balanced'. As examples of the former we have CERVOK (8 responses - but the response 's3vQk': accounts for 75.61%) and AMINGE (9 responses - but the response '{mIn_': accounts for 70.73%), and of the latter ELUCH (22 responses, no single response accounting for more than 10% of observed responses) - and across our sample these two 'types' of nonwords make up about equivalent proportions (dominant = 46%, balanced = 54%). This information is provided to illustrate that our sub-sample item set reflects the principle that patterns of responses to nonwords can vary not just with respect to how many different answers might be provoked but also that these responses can vary in their proportions of total responses, and we have tried to select a subsample that varies in both degrees.

Finally, in our subsequent analyses, we will seek to determine the proportions of singularly distinct responses generated by our individual patients. In our earlier discussion of this issue, we mentioned that our inspection of the full database provided by Pritchard and colleagues found that all participants generated some proportion of such responses. In our subset of nonword items, do we observe something similar? Interestingly, of the 1397 unique responses relating to our 201 subset, there are 739 singularly distinct responses (53%), in Fig. 1 below, we present the degree to which these are distributed over our 201 NWs. What is clear is that although most NWs tend to be associated with one or two singularly

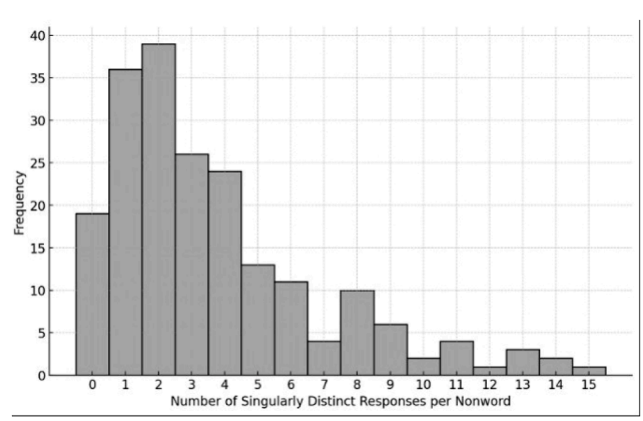

Fig. 1 – Count of singular distinct responses distribution across the 201 NWs used.

distinct responses, some are associated with a great many (UGHIME = 15, BESEALT = 14, SYCHLOM = 14).

In summary, our investigation of response patterns for our sub-set of 201 nonwords indicates that they appropriately vary in: (a) the numbers of observed different responses provoked by the 41 young healthy readers to a given nonword (response agreement), (b) the degree to which particular responses capture the total proportion of responses to an item and (c) the numbers of provoked responses that are singularly distinct being observed in the responses of only a single participant. The large variation in all three of these variables makes us confident our subset suitably represents the variation observed in the larger set of 915 nonwords. It will also be interesting to determine the degree to which these different variables may have a subsequent impact on our three approaches for classifying errors (outlined earlier). In any case, having observed these different characteristics for our nonword set, we might ask is a similar pattern of variation observed with an older adult population?

2.1.2. *Older adult reference sample*

As was discussed earlier, given we aim to test dementia patients presenting as candidates for PD, this naturally means individuals are older adults. It was therefore important to collect a sample of 'healthy' older adults reading our critical 201 NW set, which has as yet rarely been obtained. In this case, we have a database of 25 participants (12 males/13 females – mean age 67.32, SD = 3.16). All participants were recruited *via* the University of Swansea older adult participant volunteer panel. All reported English as their first language and no history of neurological disease, dementia, acquired language disorder, or developmental reading disorder. All responses produced by the cases were transcribed by a doctoral research assistant trained in phonetics – response transcriptions were also confirmed by the principal investigator and an additional academic team member who is an expert in phonetics. Our new older adult database thus comprises 5025 individual responses (25 × 201). Of these 1250 were singularly 'unique' responses. Appendix 2 provides the list of nonwords (see OSF: https://doi.org/10.17605/OSF.IO/XMZ49), with each of the different responses associated with that nonword and the relative proportion associated with these different responses. Like the observations with young participants, only 9/201 nonwords provoked a single consistent response across older participants (RA = 1). In addition, consistent with our observations with younger participants, several nonwords provoked a large range of responses. Similarly, the proportions of these responses for the nonwords varied substantially. In Appendix 2, we have sub-categorised nonwords as either 'dominant' response (70%+ of observed responses) or 'balanced' (see earlier discussion) - across our sample these two 'types' of nonwords comprise about equivalent proportions (dominant = 40%, balanced = 60%). Finally, in our earlier inspection of the younger adult database, we also consider the proportions of singularly distinct responses and of our 1250 unique responses; 682 responses were so classified (55%), a proportion similar to that observed with the younger adult database. In Fig. 2, we present the degree to which these are distributed across the 201 NWs. Again, the distribution appears to mimic that seen for young participants (see Fig. 1)

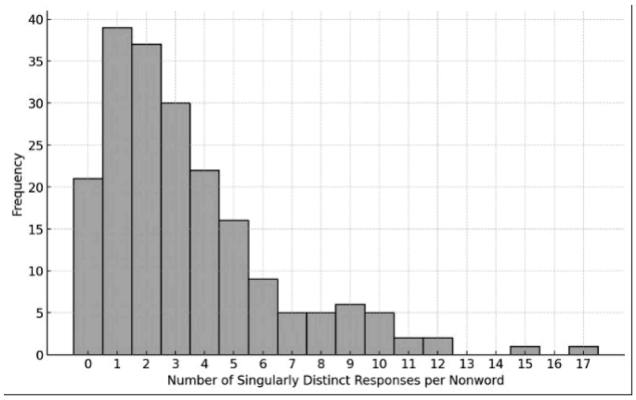

Fig. 2 – Count of singular distinct responses distribution across the 201 NWs used.

in that items can range considerably in this dimension (e.g., ERUGE = 15, SHERUCH = 17). In summary, our investigation of the response patterns in our older database suggests that, like the 41 young healthy readers, we see similar variations in item response agreement, response profiles for items, and counts of singular distinct responses linked to items. Although in several instances the key NW items observed to provoke these patterns differed across populations. In the next section, we compare older and younger health reading performance.

2.1.3. *Older vs younger adult reference samples*

Three outcomes were analysed to compare normative nonword reading performance between younger and older adult control participants: (1) the count of responses matching the modal response across the 201 item nonword set, (2) each participant's average pairwise response overlap proportion with all other participants, and (3) the count of singularly distinct responses - those generated by only a single participant. However, before moving to the patients, we are left with an important decision regarding whether to use as our normative reference sample the mean and standard deviation information provided by both samples combined or to focus just on the older adults. To investigate this issue, we undertake some comparative statistical analyses across the younger/older adult groups. For each outcome, the appropriate statistical model was selected empirically based on the outcome's data structure, distributional diagnostics, and model comparison *via* AIC, rather than applying a single analysis framework uniformly across all three – all analyses were coded using R and specific data analysis packages, and scripts are available on our OSF link.

In the first instance, the mode-match count is a bounded integer (0–201), and participants are expected to vary in their underlying probability of matching the mode as a function of individual differences in latent ability. The beta-binomial general linear model (GLM) was selected as the appropriate distributional model because each participant's score represents a bounded count—the number of nonwords out of 201 for which their response matched the modal response—and participants vary in their underlying probability of mode-matching as a function of individual differences. This

individual variation produces overdispersion relative to a pure binomial model, meaning that between participant variance in mode-match counts exceeds what a pure binomial predicts; this was confirmed prior to model fitting *via* a Pearson dispersion statistic computed from a standard binomial GLM. Model fit was formally compared against a pure binomial *via* AIC. A secondary model comparison additionally tested whether the degree of individual variability in mode-matching probability differed between age groups, by allowing the overdispersion parameter phi to vary freely across groups rather than constraining it to be equal. In the second instance, given that the average response overlap is a continuous proportion derived from pairwise comparisons across the full participant set; beta regression was evaluated but offered no improvement over a standard linear model (ΔAIC = .27 in favour of OLS), and there was no strong evidence of severe non-normality; visual inspection did not suggest major deviations. Accordingly, the simplest option of a Welch independent samples *t*-test was used for this outcome. In the third instance, the singularly distinct response count, a negative binomial GLM was used because the outcome is a count variable and an alternative Poisson model showed substantial overdispersion. This was evident in large variance-to-mean ratios ((Young: 3.75; Old: 5.84) indicating severe overdispersion relative to a Poisson distribution, and improved fit of the negative binomial relative to Poisson [$\chi^2(1) = 117.6$, $p < .001$; ΔAIC = 115.6]. In all three analyses, younger adults were specified as the reference category, consistent with their role as the typically established normative baseline. All analyses were conducted in *R*; the betabinomial used the betabin() function from the aod package, and the negative binomial used glm.nb() from the MASS package.

Approach 1 - Mode response counts - In this case, for each participant, we computed the number of responses that matched the identified modal response and the mean for each group. In general, the older adult sample had a lower mean count of responses that matched the mode (see Table 1 above). A Pearson dispersion statistic of 3.22 from the standard binomial GLM confirmed substantial overdispersion in the data, justifying the use of the beta-binomial. The appropriateness of the betabinomial over a standard binomial was further confirmed by a substantially lower AIC (beta-binomial: 523.0; binomial: 586.1; ΔAIC = 63.1) and a highly significant overdispersion parameter (phi = .010, $z = 4.01$, $p < .001$), indicating that participants' underlying mode-match probabilities varied meaningfully around the group mean, consistent with genuine individual differences latent ability. A further betabinomial model which allowed overdispersion to differ between age groups was explored but there was no evidence of further fit improvement (ΔAIC = 2.0 in favour of the singular shared overdispersion model). This was an interesting observation, since it seems although individuals clearly varied considerably in their mode-matching probability, the degree of that observed variability was largely statistically equivalent across age groups. Nonetheless, the singular model revealed that the older adult group had significantly lower odds of matching the modal response (OR = .791, 95% CI [.697, .896], $z = -3.67$, $p < .001$). In absolute terms, the model estimated that younger adults matched the modal response on 62.4% of nonwords on average (predicted mean count: 125.4 out of 201), compared to 56.8% for older adults (predicted mean count: 114.1 out of 201)—a difference of approximately 11 items across the full nonword set. Consequentially, it thus seems most prudent to use the older adult sample normative information specifically when evaluating our older adult patient sample with respect to this variable (i.e., for calculating z-scores etc.).

Table 1 – Mean number of responses that match the mode across groups.

	Young	Old
N	41	25
Mean	125.44	114.08
Std. Deviation	12.37	12.51
Minimum	94	90
Maximum	145	138

Approach 2 - Average response overlap - as mentioned earlier we sought to adopt a second approach (inspired by the work of Pritchard and colleagues) where each PD case would be compared to each individual healthy adult reader to determine the proportion of response overlap in each pairwise comparison. This would enable us to define the pattern of 'errors' observed in patients in a more 'holistic' fashion, by determining the degree to which the overall response pattern of patients diverges to a considerable degree. Put simply, for a given individual, we do a one-to-one comparison of the responses for each of the 201 nonwords and simply count the number that are identical to generate a proportion overlap. In Fig. 3, we present a heatmap of all these one-to-one comparisons. We can then calculate the average proportion overlap for each participant relative to the rest of the entire sample. This account captures the full range of responses that might occur in normal readers (rather than the single most dominant mode response). In general, the older adult sample had a lower mean proportion overlap and a larger range across this mean (see Table 2). A Welch independent samples *t*-test was conducted to compare the proportion overlap average between the two groups. The results revealed younger adults showed significantly higher average response overlap than older adults (Young: $M = .470$, $SD = .034$; Old: $M = .441$, $SD = .038$), with a mean difference of .029 (95% CI [.010, .048]), $t(45.6) = 3.13$, $p = .003$, Hedge's $g = .81$. This large effect indicates that on any given nonword, older adults' responses were, on average, approximately 3 percentage points less similar to the broader population of healthy adult responses than those of younger adults—consistent with the group difference observed under the mode-match metric, and reinforcing the conclusion that age-appropriate normative data are necessary when evaluating older adult nonword reading performance.

Approach 3 - Average number of singularly distinct responses - a third analysis approach sought to focus on the specific example of what we have dubbed 'singularly distinct' responses, defined as a response generated by only a single individual. The key point is that, if acquired phonological dyslexia disrupts the cognitive reading system, leading to a qualitative shift in reading responses, it is reasonable to assume that an increase in atypical (singularly unique)

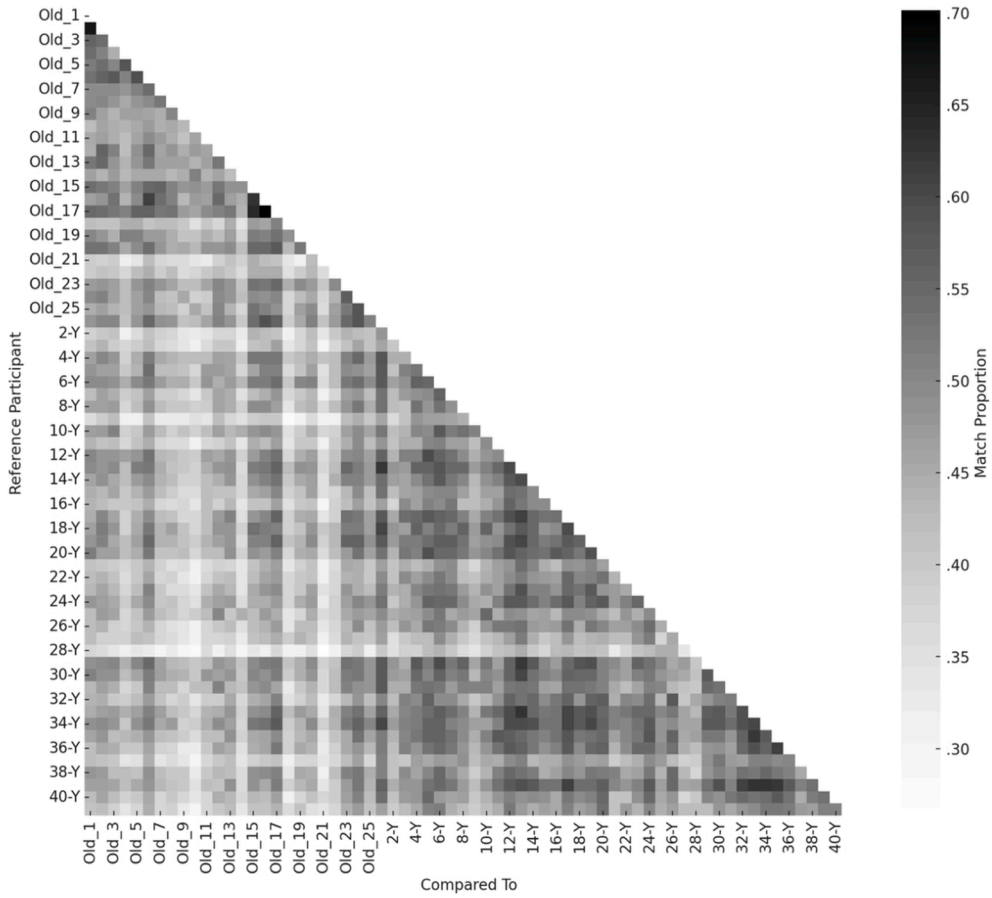

Fig. 3 – Heatmap of response overlap between individual participants' nonword response sets.

Table 2 – Mean proportion matching profile proportion across age groups

	Young	Old
Valid	41	25
Mean	.47	.44
Std. Deviation	.034	.038
Minimum	.377	.357
Maximum	.517	.504

Table 3 – Mean singularly distinct responses across age groups

	Young	Old
Valid	41	25
Mean	14.85	21.08
Std. Deviation	7.46	11.1
Minimum	4.000	6.000
Maximum	38.000	53.000

responses could indicate the severity of the damage to this system. This analytic approach was designed to reflect this assumption, with the idea that a higher number of such distinct responses may serve as a marker of more severe impairment in cases of phonological dyslexia. As in previous cases, we calculated average response counts (see Table 3 and Fig. 4), in general, the older adult sample had a higher mean count of such responses and a larger range across this mean. As mentioned earlier, a negative binomial provided a dramatically better fit than a comparative Poisson baseline [LRT: $\chi^2(1) = 117.6$, $p < .001$; $\Delta AIC = 115.6$], and the NB Pearson dispersion of 1.086 confirmed the model adequately absorbed the overdispersion. The NB also outperformed a standard linear model ($\Delta AIC = 17.8$ in favour of NB), and the estimated dispersion parameter was well-determined (theta = 5.50, SE = 1.25). Older adults produced significantly more singularly distinct responses than younger adults (IRR = 1.419, 95% CI [1.114, 1.808], $z = 2.84$, $p = .005$), with fitted mean counts of 21.08 and 14.85 respectively—meaning older adults generated approximately 42% more responses that were not observed in any other participant. Notably, this is the largest group effect of the three metrics, consistent with the likelihood that singularly distinct responding may be a particularly sensitive index of divergence from normative reading behaviour.

In summary, all three of our key measurement metrics appear to proportionally differ across older and younger adult group analyses. Interestingly, there is little work looking at the differences across the

lifespan of reading nonwords aloud, and our works suggests this clearly warrants more investigation. Moreover,

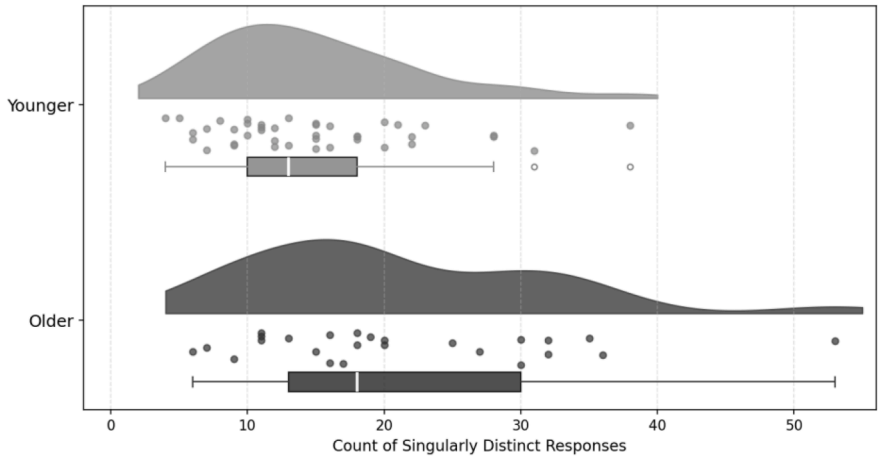

Fig. 4 – Mean singularly distinct responses for individuals in each age group

given these group differences and the fact our patient sample are all older adults, it is prudent to adopt a focus on using a mean/standard deviation from a comparative healthy population of older adults for our subsequent analyses (i.e, see Tables 1–3). Critically, if we had used such information from a normative young sample only, there was a likely risk of overestimating the degree of 'impairment' observed from the responses of PD (older adult) cases.

2.1.4. *Patient sample*

We approached local memory clinics to initially recruit patients with a neurological diagnosis of Alzheimer's disease over the period of 2015–2022 (disrupted somewhat by the global pandemic), where all individuals had expressed an interest in taking part in a research project. This resulted in a final sampled case series of 6 patients with dementia who met our criteria for potential phonological dyslexia. Potential candidates with acquired phonological dyslexia, were defined as (a) normal (often ceiling) reading on two screening tests of words and (b) 'impaired' reading with a critical set of nonwords, In this case all screening tests were taken from the PALPA battery; discussed earlier, a well-established and standardised psychometric battery used for testing patients with aphasia (Kay. Lesser & Coltheart, 1996). Specifically, the word sets were subtests PALPA 31 and 35 and the nonword subtest PALPA 36; overall accuracy scores for each patient are presented in Table 4. As was mentioned in the introduction, the PALPA nonword testing list was devised to include items with high response agreement (based on their normative screening). Thus, any responses that diverge from this putative consistent response are classified as 'errors'. In addition, 'impaired' performance reflects overall accuracy that falls below a normative 'cut-off' of two standard deviations, provided by the PALPA test battery materials. As we discussed earlier in the introduction, the PALPA manual reports that their normative sample ($N = 32$) had a mean 'accuracy' = 22.88 (SD = .72). Consequently, one can calculate that typically 4 or more 'error' responses would meet the criteria for a standard deviation 'cut-off'. Put simply, a PD suspected 'candidate' was initially defined by their making 4 or more 'errors' on the nonword (PALPA 36) list, where 'error' in this case reflects a response that deviates from the mode response provided in the handbook. On that basis, our 6 candidate cases ranged from likely 'mild' (i.e. cases 2, 4 = 4 'errors') to 'severe' (i.e., case 7 = 10 'errors'). The dementia case series had a mean age of 70.3 years at testing, range 66–78, compared with the older adult normative sample mean of 67.32 years, SD = 3.16.

2.1.5. *Performance of dementia PD patients on our critical nonword reading set*

In addition to testing patients on reading, cases were tested on a variety of cognitive assessments, including: semantic memory, episodic memory and phonological processing – this was to provide a detailed cognitive profile for each patient commensurate with a neuropsychological case series (i.e., a group of detailed case studies). We provide full details of each of these tests in Appendix 3 (see OSF: https://doi.org/10.17605/OSF.IO/XMZ49), but in every case, they are measures often used in clinical assessment for aphasia and dementia. Inspection of scores on these additional tests determined that all have profound general memory problems consistent with their diagnosis of Alzheimer's disease (see Table 4). As part of this profile, several cases presented with anomia (i.e., poor picture naming), and of these, several additionally had some observed problems with semantic memory. All cases presented with some degree of phonological impairment, which manifested as issues with nonword repetition and/or rhyme/homophone judgment - in the latter case, we have a word and a nonword version of the homophone judgement task, and naturally, all were impaired at the nonword version (in all cases we attribute impaired performance to accuracy levels below a 2 standard deviation 'cut-off'). All candidate PD case series were presented with the list of 201 nonwords from which we have earlier obtained a normative sample of healthy older adults. In addition, all elements of transcription analysis undertaken with older adults were mirrored in the patient sample. Thus, we now have two new databases of transcribed responses for 201 items across 6 PD cases and 25 older healthy adults. We provide this unique dataset for online use at the following website (see OSF: https://doi.org/10.17605/OSF.IO/

Table 4 – Background neuropsychological screening for case series.

	ID	Case 6	Case 1	Case 5	Case 3	Case 4	Case 2
Age		66	74	78	68	67	69
Handedness		Right	Right	Left	Right	Right	Right
General memory	**Cutoff**						
ACE	88	48/100	73/100	77/100	71/100	68/100	63/100
Hopkins verbal mem	24.5	2-2-3 (7)	4-4-4 (12)	2-4-4 (10)	4-4-4 (12)	3-2-4 (9)	2-4-5 (11)
Rey Figure	22	Const apraxia	28/36	34/36	22/36	32/36	Const apraxia
Semantics							
PPT	47	Vis agnosia	52/52	43/52	52/52	52/52	48/52
Writ synonyms - PALPA	55	55/60	59/60	50/60	57/60	60/60	53/60
Naming - PALPA	38	30/40	40/40	37/40	40/40	30/40	30/40
Phonology - PALPA							
Repetition – words	38	39/40	40/40	40/40	40/40	40/40	40/40
Repetition – nwords	36	34/40	40/40	40/40	40/40	40/40	40/40
Rhyme judgment - PALPA							
Aud	54	54/60	60/60	55/60	45/60	60/60	58/60
Writ	53	50/60	54/60	41/60	40/60	57/60	44/60
Homophone judgement - PALPA							
Regular words	18.48	12/20	19/20	18/20	16/20	18/20	14/20
Irregular words	18.48	13/20	20/20	19/20	16/20	19/20	13/20
Nonwords	17.86	8/20	12/20	16/20	15/20	14/20	9/20
Reading - PALPA							
Words (Regularity)	55	60/60	58/60	60/60	56/60	60/60	56/60
Nonwords	22	14/24	15/24	19/24	18/24	20/24	20/24
Words (Freq x Imag)	76	80/80	79/80	80/80	78/80	80/80	78/80
Lexical decision	115.24	–	117/120	116/120	117/120	119/120	116/120

**Const Apraxia = patient presents with constructional apraxia, thus unable to do task. Vis agnosia = Visual agnosia unable to do task.

XMZ49), for use by future researchers interested in computational modeling of acquired phonological dyslexia (e.g., Nickels et al., 2008).

In the introduction, we highlighted three possible means to operationalise 'accuracy' for nonword responses, which we aimed to contrast in our testing of the patients as PD, and thus characterise the severity of their impairment. Firstly, we can calculate the count of their responses that match the mode response for a given nonword as defined by the older adult normative sample. This is a typical approach and how we initially screened individuals using the PALPA36 nonword reading list (see above). Secondly, we can calculate the proportion of response overlap the patient has with each of the control participants and compare it to the typically observed average proportion overlap. Thirdly, we count the number of singularly distinct responses the patient generates and compare that to the average seen in the normative sample. We present each of these approaches in separate sections below.

2.1.6. *Performance of dementia PD cases as a function of mode response 'accuracy'*

If we adopt the straightforward approach of identifying the 'mode' responses for each of our 201 nonwords (as determined by the older adult normative sample) and then scoring individual patient responses as 'correct' or not based on whether they match with these target responses, this will yield a count for each patient. We can then compare these counts to the mean and standard deviation information provided by the older adult controls normative sample (see Table 1 - Mean: 114.08, SD: 12.51). Given that we have a case series of patients, we sought to determine the severity of impairment for each individual case study rather than combining them into a single group. In this case, we can then calculate the severity of impairment using the specialist software provided by Crawford, Garthwaite & Porter (2010 - Singlims_ES.exe). Given that the program tests whether an individual's score is significantly different from a normative sample, it is ideally designed for single-case neuropsychological work. The software provides a point estimate of the effect size for the difference between the case and controls (Z-CC) with an accompanying 95% confidence interval. We can also calculate a point and interval estimate of the abnormality of the case's score (i.e., an estimate of the percentage of the population that would obtain a lower score). On this basis, it is apparent that 4/6 cases are generating counts of modal responses that are significantly lower than the normative sample (see Table 5), with a clear range in severity of impairment (i.e., case 3 is borderline). Meanwhile, two other cases (2 and 4) performed within a 'normal' range; it is of note that these same cases presented with the 'mildest' level of error profile based on PALPA 36 remain unimpaired (see Table 4).

2.1.7. *Performance of dementia PD cases as a function of the proportion of response overlap with individual healthy older adult controls*

A second approach was to consider the overall response profile of each of our PD cases to each of our individual older adult controls. That is, calculate an average proportion response overlap. We can then calculate a z-score for each patient

Table 5 – Single case studies - Mode count impairment profile.

Patient	Mode count	t-value	*p*-value	Effect size (Z-CC)	95% CI	Pop % estimate
Case 6	41	−5.73	$p = .00001$	−5.84	(-7.52, −4.15)	.0003%
Case 1	65	−3.85	$p = .0008$	−3.92	(-5.09, −2.75)	.039%
Case 5	78	−2.83	$p = .009$	−2.88	(-3.79, −2.75)	.46%
Case 3	89	−1.97	$p = .03$	−2.01	(-2.68, −1.31)	3.05%
Case 2	100	−1.10	n.s	−1.12	(-1.62, −.61)	14.03%
Case 4	104	−.79	n.s	−.81	(-1.25, −.35)	21.86%

Table 6 – Single case studies - Average proportion overlap impairment profile.

Patient	Average overlap	t-value	*p*-value	Effect size (Z-CC)	95% CI	Pop % estimate
Case 6	.17	−6.97	$p = .00001$	−7.11	(-9.13, −5.07)	.00002%
Case 1	.26	−4.64	$p = .00005$	−4.74	(-6.12, −3.34)	.005%
Case 5	.295	−3.74	$p = .0005$	−3.82	(-4.95, −2.67)	.05%
Case 3	.36	−2.06	$p = .05$	−2.11	(-2.81, −1.39)	2.50%
Case 4	.38	−1.55	n.s	−1.58	(-2.16, −.98)	6.73%
Case 2	.4	−1.03	n.s	−1.05	(-1.54, −.55)	15.61%

based on the mean and standard deviation provided by the normative sample (see Table 2 - Mean = .44, SD = .038). Again, these are presented as a case series of patients (see Table 6), and as before, the average similarity overlap is statistically compared to the normative sample using the approach outlined by Crawford (computing effect sizes with confidence intervals and population percentage estimates). Overall, we appear to observe a similar pattern of classification of the cases; with the same 3 cases being severely impaired, 1 case being borderline, and 2 cases within normal range. Similarly, the order of severity of impairment for the 3 impaired cases appears identical. Interestingly, the general effect sizes appear to be larger than those observed under the mode count approach (compare Tables 5 and 6).

2.1.8. *Performance of dementia PD cases as a function of the count of exclusively singularly distinct responses*

A third approach was to count the number of the singular distinct count responses for each of our PD cases under an identical approach to the previous approaches. What was striking in that analysis was just how many such responses are typically observed in our older adult population (see Table 3 - Mean = 21.08, SD = 11.1), with on average of 10% of responses for a given participants being so classified (a proportion similar to that observed earlier in the Pritchard work). It appears thus that there are many unique examples from controls where the phonetic variation in their vowel qualities produce 'non-typical' responses but would be understood because they are in range of nearby vowels. And, within the diversity of English language variation worldwide, would be mutually recognisable to most speakers. For example, variation between front high vowels [I] & [i], front low vowels [E] & [{] and central vowels [@], [3] & [V] have some variation in the control dataset. A lay listener may not hear these as being all that 'distinct' given that people tend to 'allophonically' tolerate several 'phones' (i.e. sound units) for any one phoneme (perception of that sound), but phonetically they are indeed unique and so may contribute to this wide range of responses given by controls. For example, [pr@gis] for PROGESE has a higher front vowel in the second syllable; another, [rEsiZ] for RASIGE has a higher mid vowel in the first syllable than most. In any case, we can then compare these average singularly unique counts to the mean and standard deviation information provided for the older adult controls to calculate the severity of impairment using the approach outlined by Crawford - and as before we can calculate effect sizes with confidence intervals and population percentage estimates. Interestingly, in this case, whilst the same three key candidates identified earlier as being severely impaired, remain so under this approach (with the same ranked order of severity) - the case earlier identified as 'borderline' (Case 3) is now clearly significantly impaired.

Moreover, one additional patient (Case 2) is now apparently borderline. This is also consistent with the fact that this approach also generally yields the highest effect sizes (Z-CC) of any of the three approaches (see Tables 5–7).

Overall, our three approaches have converged on a similar picture of the severity of PD nonword reading impairment. However, the third approach yields a picture of generally higher severity 'profiles' across the board for all cases. In Table 8, we provide a summary of the findings with the cases largely ranked by severity of impairment. Interestingly, three of the cases failed all the semantic tests (Case 6, 1 and 5) and two passed them all. But there is no evidence that these patterns of impairment are related to the severity of observed nonword reading impairment. Similarly, the cases appear to vary concerning phonological test impairments, Case 6 passing only one test and Case 1 only failing the nonword homophone judgement task. But again, there is no obvious evidence that these patterns of impairment are related to the severity of observed nonword reading impairment 'profile'. Overall, there was no clear evidence that the degree of disruption to our tasks aiming to index either semantic or phonological impairment was obviously linked to the severity of PD presentation as classified by our three different classification approaches. Suggesting that the problems occurring in patients with nonword reading are likely a manifestation of a deterioration of a reading-specific cognitive system (see Tree, 2008).

Table 7 – Single case studies - Singularly unique response count impairment profile.

Patient	Unique count	t-value	p-value	Effect size (Z-CC)	95% CI	Pop % estimate
Case 6	129	9.53	$p = .0000011$	9.72	(6.96, 12.48)	.000001%
Case 1	83	5.47	$p = .000001$	5.58	(3.96, 7.19)	.0006%
Case 5	67	4.06	$p = .0005$	4.14	(2.90, 5.36)	.02%
Case 3	62	3.61	$p = .001$	3.69	(2.57, 4.79)	.07%
Case 2	44	2.02	$p = .05$	2.06	(1.36, 2.76)	2.70%
Case 4	37	1.41	$p = .17$	1.43	(.86, 1.99)	8.60%

Table 8 – Single case studies - Singularly unique response count impairment profile.

Case	Semantics tests	Phonology tests	Mode Z-CC	Av profile Z-CC	Sing unique Z-CC
Case 6	0/3	1/7	−5.84	−7.11	9.72
Case 1	3/3	6/7	−3.92	−4.74	5.58
Case 5	0/3	5/7	−2.88	−3.82	4.14
Case 3	3/3	2/7	−2.01	−2.11	3.69
Case 2	0/3	3/7	−1.12	−1.05	2.06
Case 4	2/3	6/7	−.81	−1.58	1.43

3. General discussion

The primary aim of this study was to evaluate how best to classify the severity of nonword reading impairment in older adults presenting with acquired phonological dyslexia (PD). To do this, we compared three approaches to scoring patient responses against normative nonword reading data: (a) a moderesponse accuracy approach (the method most typically used), (b) an average response-overlap approach (inspired by Pritchard et al., 2012), and (c) a singularly distinct response count approach, indexing responses produced by only one individual relative to the control response set. A central contribution of the present work is the demonstration that normative performance differs by age group. Across all three metrics, older adults differed reliably from younger adults. This has an immediate methodological implication: when evaluating older patients for PD, age-appropriate normative data are not optional but essential. Had we relied on younger adult norms alone, the severity of impairment in our patient sample would likely have been overstated. More broadly, these findings highlight an underexplored issue in the literature: the effects of healthy ageing on nonword reading variability itself.

A second contribution concerns the nature of variability in "normal" nonword reading. Across both younger and older controls, participants routinely produced responses that were singularly distinct (i.e., responses not shared by any other participant for that item). This is important because it challenges any simple assumption that atypical-looking responses are necessarily pathological. In our data, singularly distinct responses were not rare edge cases; they were a regular feature of healthy performance. This has clear consequences for how 'errors' should be conceptualised in nonword reading, and it reinforces the difficulty of classifying patient responses in the absence of large and appropriate normative datasets. At the same time, the presence of singularly distinct responses in controls does not make the construct uninformative. Rather, it shifts the emphasis from presence/absence to degree, consistent with impairment being on a quantitative continuum. Our third approach (counting singularly distinct responses) appears to capture this distinction effectively: although all three approaches broadly converged on a similar rank ordering of patient severity, the singularly distinct response metric yielded the strongest separation from controls and identified a broader range of impairment severity (including one case that appeared only borderline under the other methods). This suggests that the metric may be particularly sensitive to qualitative disruption in nonword reading performance, and we would advise its inclusion in further studies of PD presentation in future.

Notably, the three scoring approaches produced a broadly similar overall picture of the case series, with consistent identification of the most severely impaired cases and similar rank ordering of severity. This convergence is reassuring and supports the validity of all three approaches as tools for characterising PD. However, the approaches are not interchangeable. The mode-response method provides a familiar and clinically interpretable summary, the response-overlap method better captures the full distribution of normal variation, and the singularly distinct response method appears especially sensitive to atypicality beyond the observed control response space. Taken together, these approaches provide a richer framework for describing nonword reading impairment than any single metric alone. The present findings also raise questions for computational accounts of reading aloud. Current models are typically evaluated against aggregate or modal human responses, but our data highlight substantial individual-level variability, including healthy singularly distinct responses. If such variability is a stable property of skilled reading, then reading models likely need to account for a broader range of plausible outputs, not just the dominant response. In this respect, our findings extend the observations of Pritchard and colleagues and suggest that existing reading model evaluations may underestimate the breadth of normal human variation in nonword pronunciation.

Finally, although our patients varied in the severity of semantic and phonological impairments on auxiliary neuropsychological measures, we found no obvious one-to-one relationship between these broader profiles and the severity of nonword reading impairment as classified by our three scoring methods. This is consistent with the view that PD can reflect disruption within reading-related processes that is not straightforwardly reducible to general semantic or phonological impairment severity alone (cf. Tree, 2008). That conclusion should be treated cautiously given the small case-series design, but it reinforces the value of detailed, response-level analysis when characterising acquired dyslexic syndromes.

3.1. *Item-level implications and test development*

Although the present study focused primarily on participant-level classification, we undertook a further exploratory item-level analysis to determine whether some nonwords may be more diagnostically informative than others. Specifically, items with high control consensus (strong modal responding) but frequent patient novelty (responses not observed in controls) may offer particularly strong patient–control separation. Thus, for each item, we first quantified control consensus as the proportion of control participants producing the modal phonetic transcription (consistent with our first analysis approach). We then quantified patient divergence as the proportion of patient transcriptions that were novel with respect to controls for that same item; operationalised as transcriptions not observed anywhere in the control responses for that nonword (consistent with approach 3). These two components were combined into a single heuristic 'diagnostic score' computed as control_mode_prop × patient_novel_rate, such that items score highly when controls show strong agreement on a single transcription while patients frequently produce responses outside the control response set. Items were then ranked by this score to highlight candidate nonwords with maximal patient–control separation. The approach then allowed us to identify a candidate short-list of 15 nonwords for which our 'diagnostic scores' were high (see Table 9). This "bottom-up" approach is promising because a shorter, carefully selected item set could make assessment more practical in clinical contexts, where administration and transcription burden are important constraints. However, these item-level findings remain preliminary. Our ranking procedure is heuristic, and the current sample size (especially the patient sample) is likely too small to support definitive claims about diagnostic sensitivity or specificity at the item level. Replication in larger samples will be essential before such an item set could be recommended for clinical use, but the example provided here provides a useful demonstration of utility. In any case, these investigations may well suggest that a smaller testing set might act as a very useful practical proxy for clinical work as regards 'diagnosing' phonological dyslexia, and we intend to explore this in future.

3.2. *Future directions*

The present work has provided the impetus for several new and exciting directions of research. Firstly, given we have established that PD presentation can manifest in a typical AD neurodegenerative presentation (see also Passafiume et al., 2000) - it follows that it may be interesting to track the degree to which NW reading impairment increases over time and if so whether this reflects simply an even more exaggerated pattern seen at present, or whether new aspects of impairment emerge. Such longitudinal work would also further determine whether even over small-time frames idiosyncratic responses observed within patients are consistent. A related and clinically important question is whether response-level profiling of nonword reading could contribute to early differential diagnosis across neurodegenerative syndromes (e.g., Parkinson's disease, Cortico-basal degeneration) for whom acquired phonological dyslexic patterns of

Table 9 – Top 15 candidate diagnostic nonwords (controls *vs* patients).

Nonword	Control modal transcription	Control mode proportion	# Unique control transcriptions	# Patients novel *vs* controls (of 6)	Patient novelty rate	Diagnostic score
Flizzack	flIz{k	.908	4	5	.833	.756
Pazim	p{zIm	.742	6	5	.833	.619
Conglire	kQNgl7	.609	14	6	1.000	.609
Disbads	dIsb{dz	.894	4	4	.667	.596
Fizack	fIz{k	.864	6	4	.667	.576
Terbip	t3bIp	.848	7	4	.667	.566
Ulheep	Vlhip	.848	6	4	.667	.566
Adanx	{d{Nks	.621	11	5	.833	.518
Outsmaw	6tsm9	.758	7	4	.667	.505
Nultril	nVltrIl	.754	13	4	.667	.503
Noffle	nQf@l	1.000	1	3	.500	.500
Flisple	flIsp@l	.723	12	4	.667	.482
Grizzack	grIz{k	.939	4	3	.500	.470
Ruzzle	rVz@l	.939	4	3	.500	.470
Imbroths	ImbrQTs	.939	5	3	.500	.470

Columns: control mode proportion = proportion of controls producing the modal transcription; unique control transcriptions = number of distinct control pronunciations; patient novelty rate = proportion of patients whose transcription is not found in the control set; diagnostic score = control mode proportion × patient novelty rate.

performance have been reported (Ball et al., 2004; Code et al., 2006; Elorriaga-Santiago et al., 2013; Tree & Kay, 2008, 2015). If different diseases produce distinct patterns of mode deviation, response overlap, and singularly distinct responding, then nonword reading may offer a low-cost and theoretically informative behavioural marker. This remains speculative at present, but the current study provides a methodological foundation for testing it. Finally, our findings also underscore the need for larger normative datasets, particularly for older adults – and the present work provides a first available resource in this respect. The observation that singularly distinct responses occur in all participants may partly reflect genuine human variability, but it is also sensitive to sample size: as normative databases grow, some responses currently classified as 'singularly distinct' may no longer be unique. Future work should therefore examine the stability of these metrics as a function of normative sample size and demographic composition.

3.3. *Final conclusions*

This study has explored a longstanding conundrum in testing acquired phonological dyslexia: how to classify nonword reading 'impairment' when nonword pronunciation is inherently variable in healthy readers? We compared three scoring approaches—mode-response accuracy, average response overlap, and singularly distinct response count—and found that they broadly converged in their characterisation of patient severity while differing in sensitivity. As best practice we would advise using all in PD research rather than merely relying on the modal response approach. The current work has also yielded some further interesting observations including: (1) healthy nonword reading is likely more variable than is often assumed, including regular production of 'singularly distinct' responses, and (2) age-appropriate normative data are essential when assessing older adults for PD, demonstrating that age may play some important role in how nonword performance occurs. Together, these results support the use of response-level normative frameworks for PD classification and provide practical tools for future neuropsychological and computational work on nonword reading. More broadly, this approach opens new avenues for research into the cognitive and clinical heterogeneity of acquired phonological dyslexia, including longitudinal progression and potential differential diagnosis across neurodegenerative disorders.

Scientific transparency statement

DATA: All raw and processed data supporting this research are publicly available: https://osf.io/xmz49/overview?view_only=56490b8ab7064f2e8b3223f5d6b032c0.

CODE: All analysis code supporting this research is publicly available: https://osf.io/xmz49/overview?view_only=56490b8ab7064f2e8b3223f5d6b032c0.

MATERIALS: Some study materials supporting this research are publicly available, while some are subject to restrictions: https://osf.io/xmz49/overview?view_only=56490b8ab7064f2e8b3223f5d6b032c0; Materials are contained in the manuscript or supplemental files.

DESIGN: This article reports, for all studies, how the author(s) determined all sample sizes, all data exclusions, all data inclusion and exclusion criteria, and whether inclusion and exclusion criteria were established prior to data analysis.

PRE-REGISTRATION: No part of the study procedures was pre-registered in a time-stamped, institutional registry prior to the research being conducted. No part of the analysis plans was pre-registered in a time-stamped, institutional registry prior to the research being conducted.

For full details, see the *Scientific Transparency Report* in the supplementary data to the online version of this article.

Author contributions

Conceptualization: JT.; Methodology: J.T, BJ.; Formal Analysis: J.T, AJ, BJ, VZ.; Writing—Original Draft Preparation: J.T, BJ.; Writing—Review & Editing: J.T, AJ, BJ, VZ.; Visualization: J.T.

Informed consent statement

Informed consent was obtained from all participants involved in the study.

Declaration of generative AI

AI and AI-assisted technologies in the manuscript preparation process During the preparation of this work the author (JT) used Claude AI in order to act as a potential external reviewer for a draft of the manuscript and check/edit R code and check datafiles. After using this tool/service, the author (JT) reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding

This research was partly supported by funding from the British Academy small grant scheme - SG151776 Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by an IRAS Ethics committee (Project ID: 234,933).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

The authors would like to thank all the participants for taking part in this study and for sharing their experiences. Their collaboration and support were invaluable to this research.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2026.09.012.

REFERENCES

Andrews, S., & Scarratt, D. R. (1998). Rule and analogy mechanisms in reading nonwords: Hough dou peapel rede gnew wirds? *Journal of Experimental Psychology: Human Perception and Performance, 24*(4), Article 1052.

Ball, M. J., Code, C., Tree, J., Dawe, K., & Kay, J. (2004). Phonetic and phonological analysis of progressive speech degeneration: A case study. *Clinical Linguistics & Phonetics, 18*(6–8), 447–462.

Beauvois, M. F., & Derouesne, J. (1979). Phonological alexia: Three dissociations. *Journal of Neurology, Neurosurgery, and Psychiatry, 42*(12), 1115–1124.

Code, C., Muller, N., Tree, J., & Ball, M. (2006). Syntactic impairments can emerge later: Progressive agrammatic agraphia and syntactic comprehension impairment. *Aphasiology, 20*(9), 1035–1058.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204–256.

Coltheart, M., Saunders, S. J., & Tree, J. J. (2010). Computational modelling of the effects of semantic dementia on visual word recognition. *Cognitive Neuropsychology, 27*(2), 101–114.

Coltheart, M., & Ulicheva, A. (2018). Why is nonword reading so variable in adult skilled readers? *PeerJ, 6*, Article e4879.

Elorriaga-Santiago, S., et al. (2013). Phonological processing in Parkinson's disease: A neuropsychological assessment. *NeuroReport, 24*(15), 852–855.

Friedman, R. B. (1995). Two types of phonological alexia. *Cortex, 31*(2), 397–403.

Funnell, E. (1983). Phonological processes in reading: New evidence from acquired dyslexia. *British Journal of psychology, 74*(2), 159–180.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology. Human Perception and Performance, 5*(4), Article 674.

Kay, J., Lesser, R., & Coltheart, M. (1996). Psycholinguistic assessments of language processing in aphasia (PALPA): An introduction. *Aphasiology, 10*(2), 159–180.

Ktori, M., Tree, J. J., Mousikou, P., Coltheart, M., & Rastle, K. (2016). Prefixes repel stress in reading aloud: Evidence from surface dyslexia. *Cortex, 74*, 191–205.

Masterson, J., Hazan, V., & Wijayatilake, L. (1995). Phonemic processing problems in developmental phonological dyslexia. *Cognitive Neuropsychology, 12*(3), 233–259.

Mousikou, P., Sadat, J., Lucas, R., & Rastle, K. (2017). Moving beyond the monosyllable in models of skilled reading: Mega-study of disyllabic nonword reading. *Journal of Memory and Language, 93*, Article 169192.

Nickels, L., Biedermann, B., Coltheart, M., Saunders, S., & Tree, J. J. (2008). Computational modelling of phonological dyslexia: How does the DRC model fare? *Cognitive Neuropsychology, 25*(2), 165–193.

Passafiume, D., Di Giacomo, D., & Giubilei, F. (2000). Reading latency of words and nonwords in Alzheimer's patients. *Cortex, 36*(2), 293–298.

Pritchard, S. C., Coltheart, M., Palethorpe, S., & Castles, A. (2012). Nonword reading: Comparing dualroute cascaded and connectionist dual-process models with human data. *Journal of Experimental Psychology: Human Perception and Performance, 38*(5), Article 1268.

Sartori, G., Barry, C., & Job, R. (1984). *Phonological dyslexia: A review. Dyslexia: A global issue*, Article 339356.

Seidenberg, M. S., Plaut, D. C., Petersen, A. S., McClelland, J. L., & McRae, K. (1994). Nonword pronunciation and models of word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 20*(6), Article 1177.

Tree, J. J. (2008). Two types of phonological dyslexia–A contemporary review. *Cortex, 44*(6), 698–706.

Tree, J. J., & Kay, J. (2006). Phonological dyslexia and phonological impairment: An exception to the rule? *Neuropsychologia, 44*(14), 2861–2873.

Tree, J. J., & Kay, J. (2008). Longitudinal assessment of language and memory impairments in pathologically confirmed cortico-basal ganglionic degeneration. *Cortex, 44*(9), 1234–1247.

Tree, J., & Kay, J. (2015). Longitudinal assessment of short-term memory deterioration in a logopenic variant primary progressive aphasia with post-mortem confirmed Alzheimer's disease pathology. *Journal of Neuropsychology, 9*(2), 184–202.

Tree, J. J., & Playfoot, D. (2015). Declining object recognition performance in semantic dementia: A case for stored visual object representations. *Cognitive Neuropsychology, 32*(7–8), 412–426.