Establishing and evaluating the gradient of item naming difficulty in post-stroke aphasia and semantic dementia

Erling Nørkær[1], Ajay D. Halai[2], Anna Woollams[3], Matthew A. Lambon Ralph[2†*], Rahel Schumacher[2,4†*]

[†] these authors share senior authorship

* corresponding authors

[1] Department of Psychology, University of Copenhagen, Copenhagen, Denmark

[2] MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, United Kingdom

[3] Division of Neuroscience and Experimental Psychology, University of Manchester, Manchester, United Kingdom

[4] Neurocenter, Luzerner Kantonsspital, Lucerne, Switzerland

Corresponding addresses:

MRC Cognition and Brain Sciences Unit, University of Cambridge, 15 Chaucer Road, Cambridge, CB2 7EF, United Kingdom

matt.lambon-ralph@mrc-cbu.cam.ac.uk

rahel.schumacher@luks.ch

Abstract

Anomia is a common consequence following brain damage and a central symptom in semantic dementia and post-stroke aphasia, for instance. Picture naming tests are often used in clinical assessments and experience suggests that items vary systematically in their difficulty. Despite clinical intuitions and theoretical accounts, however, the existence and determinants of such a naming difficulty gradient remain to be empirically established and evaluated. Seizing the unique opportunity of two large-scale datasets of semantic dementia and post-stroke aphasia patients assessed with the same picture naming test, we applied an Item Response Theory (IRT) approach and we (a) established that an item naming difficulty gradient exists, which (b) partly differs between patient groups, and is (c) related in part to a limited number of psycholinguistic properties - frequency and familiarity for SD, frequency and word length for PSA. Our findings offer exciting future avenues for new, adaptive, time-efficient, and patient-tailored approaches to naming assessment and therapy.

Keywords: *Naming, Item Response Theory, Post-stroke Aphasia, Semantic Dementia*

1. Introduction

Anomia – problems in naming objects and concepts – is a common consequence following brain damage of numerous aetiologies. It is one of the central symptoms in semantic dementia and post-stroke aphasia, for instance (Hodges & Patterson, 2007; Kohn & Goodglass, 1985; Lambon Ralph et al., 2001; Woollams et al., 2008). Hence, naming tests are often part of a clinical evaluation in various patient groups. Clinical experience strongly suggests that there are systematic variations across items, making certain items harder to name than others. These systematic differences in item naming difficulty are also implied in classical models of naming, e.g. reflected as different activation thresholds in Morton's (1969) logogen model or in Dell and colleagues' (1997) model of lexical access in picture naming. Despite strong clinical intuitions and theoretical accounts, however, the existence and determinants of such a naming difficulty gradient remain to be empirically established and evaluated. Seizing the unique opportunity of two large-scale datasets of semantic dementia and post-stroke aphasia patients assessed with the same picture naming test, we applied an Item Response Theory (IRT) approach to explore: (a) if there are systematic differences in item naming difficulty, (b) if the naming difficulty gradient is different across aetiologies, and (c) how much these item difficulty gradients relate to psycholinguistic word properties. Establishing the presence and nature of a systematic gradient of item naming difficulty would allow for new approaches to systematic test construction and adaptive, time-efficient assessments tailored as needed for each patient group, as well as unlock new approaches to naming therapy that utilize the difficulty gradients to create “zone of proximal re-development” rehabilitation programmes (Conroy et al., 2012).

1.1 Establishing a systematic gradient of item naming difficulty

Many different naming tests exist and some were developed to contain items of varying difficulty. Among the most widely used examples are the Graded Naming Test (GNT; (Mckenna & Warrington,

1983)) and the Boston Naming Test (BNT; (Goodglass et al., 1983)) in English, and the naming test of the Aachener Aphasie Test (AAT; Huber et al., 1983) in German. However, little research has evaluated whether the variation of item difficulty aligns with patients' naming success and if this is the same across groups.

Given a sufficiently large and appropriate dataset, one formal approach to establishing the presence and nature of a difficulty gradient in patients' performance is IRT. IRT provides a powerful statistical framework for deriving psychometric properties of individual items (Thomas, 2019). The core principle in IRT is to derive a mathematical model that estimates one or more parameters pertaining to each item using individual item response results (Rasch, 1960). By adopting an IRT approach, we can ask whether items vary in naming difficulty, and whether the items discriminate well between individuals across different levels of anomia severity.

One of the few previous studies employed IRT to investigate the item parameters of the BNT in a sample of 300 (non-aphasic) patients (Pedraza et al., 2011). Similarly, in a study of 69 patients with mild Alzheimer's disease, Graves et al. (2004) adopted an IRT approach to derive item parameters for the BNT with the aim of comparing different short versions of the test. While these studies do address the question of whether an item difficulty gradient exist, they were performed with participants who did not (at least not necessarily) present with naming problems and therefore cover only a limited range of possible anomia severity.

1.2 The influence of aetiology on the item naming difficulty gradient

If item naming difficulty gradients do exist, it is relevant to evaluate if they are the same across patient groups. The importance of the question is twofold. First, it is important to determine whether it is reasonable to model item parameters so that they are freely estimated across patients with different aetiologies. This is only warranted if it can safely be assumed that items are equally difficult for these different patient groups. Second, *if* some items are more difficult for one patient

group than another regardless of the level of anomia severity it may suggest the need for tailored assessment tools for different patient cohorts. One way to investigate whether item difficulty differs between aetiologies is by means of analysis of differential item functioning (DIF), an IRT based method that assesses whether the difficulty (and/or discrimination) gradient varies between groups (Chalmers et al., 2016; Teresi et al., 2021).

1.3 Relating the difficulty gradient to item properties

If a systematic gradient of item naming difficulty can be established, it is of interest to explore how much of the item difficulty gradient is related to different factors (e.g. psycholinguistic properties). Furthermore, if the gradient appears to be different between aetiologies, this might also be reflected in a different pattern of relevant factors.

Past work has tackled elements of this two-part question. An older literature looked directly at the relationship of item properties to naming success in post-stroke aphasia (Ellis et al., 1996; Nickels & Howard, 1995) and semantic dementia (Lambon Ralph et al., 1998) but did not establish the gradient of item naming difficulty itself. In more recent research, Fergadiotis et al. (2015) used IRT to establish a naming gradient in PSA on the Philadelphia Naming Test and related this to psycholinguistic item properties, but did not explore how this 'psycholinguistic makeup' differed between aetiologies. Thus, to our knowledge, the full two-part exploration (establishing the relationship between a gradient of item naming difficulty and (psycholinguistic) item properties, and then exploring how it differs between aetiologies) remains to be achieved. The answer is important not only for advancing the understanding of the bases of naming impairments, but also because it potentially unlocks new approaches to naming assessment and therapy based on effective and efficient sampling of the relevant item properties.

2. Methods

2.1 Participants

This study analysed data from two large samples of patients with two different aetiologies: a sample of 80 patients with chronic post-stroke aphasia (PSA), reported for instance in Halai et al. (2020) plus some new cases recruited with the same inclusion criteria, and a sample of 60 patients diagnosed with semantic dementia (SD) (Woollams et al., 2008) who were assessed longitudinally (yielding a total of 160 observations). Given the significant decline between testing sessions, the longitudinal SD data were treated as independent observations in the analyses, in line with previous publications (Woollams et al., 2008). Table 1 contains characteristics of each sample. Informed consent was obtained from all participants prior to participation, in line with the Declaration of Helsinki and as approved by the local NHS ethics committee.

Table 1. *Characteristics of the two samples.*

			CSB Naming Performance			
Patient group	N	Mean age	Mean	SD	Min	Max
Semantic Dementia	160	62.8	25.44	19.91	0	64
Post-Stroke Aphasia	80	65.2	32.17	21.25	0	61

CSB: Cambridge Semantic Memory Test Battery, N: Number of patients/observations in that sample.

2.2 Measures

All patients were administered the naming test of the Cambridge Semantic Memory Test Battery (CSB), a set of tests used to assess semantic knowledge across different modalities in a clinical setting (Adlam et al., 2010; Bozeat et al., 2000). The naming test contains 64 black and white line drawings of common objects selected to cover 8 different semantic categories with 8 items each: domestic and foreign animals, birds, fruits, small and large household items, vehicles and tools. The object drawings were taken from the Snodgrass & Vanderwart (1980) 260 standardized picture set.

The performance of both patient groups spanned not only the full range of possible scores, as shown in Table 1, but was also evenly distributed within both groups.

2.3 Statistical analyses

2.3.1 Item analyses

IRT is a powerful tool in estimating psychometric properties of clinical assessments (Thomas, 2019). To assess the item parameters of the CSB naming test, a set of unidimensional two parameter logistic models were fitted to the dichotomous response data (naming success yes/no) of the two patient samples. All IRT modelling was conducted using the mirt R package (Chalmers, 2012). The modelling resulted in the derivation of two distinct parameters for each item. Figure 1A illustrates the item parameters of *helicopter*. The item *difficulty parameter* indicates the amount of the underlying trait, theta, required for a 50 % probability of correctly naming that object. The item *discrimination parameter* is the slope of the tangent at the point of inflection of the item characteristic curve and indicates how well an item discriminates between individuals at that level of theta. Individual person parameters (theta values for each patient) can be easily derived from the item parameters and individual patient test performance. As the IRT model is essentially a unidimensional confirmatory factor analysis, the theta values are simply individual factor scores on the extracted factor.

Initially, a Constrained Model was fitted with item parameters restricted to be equal across the two patient samples. If one expected the item parameters to be the same regardless of aetiology, this initial model could be sufficient. However, as one goal of this study was to investigate how items differed in difficulty between patient populations, the Constrained Model was used as a baseline for further analyses. From the Constrained Model, a set of anchor items for the subsequent DIF analyses was determined based on a procedure outlined by Meade & Wright (2012). Potential anchors are identified by assessing each item for potential DIF by allowing the target item's parameters to vary freely while keeping all other items constrained as anchor items (cf. Kopf et al. (2015) for different

strategies in anchor item selection). From the resulting set of identified non-DIF items, the five items with the highest discrimination parameter values were chosen as anchor items for fitting an Anchored Model. The five items were *pineapple, saw, scissors, swan* and *candle.* In the Anchored Model, these five items were constrained to have equal parameters across samples, and DIF for the remaining 59 items was investigated. Lastly, a Final Model was fitted using all non-DIF items from the preceding DIF analysis as anchor items. The rationale for this Final Model was that if items did not show substantial DIF, their parameters should not be allowed to vary freely between patient groups. Therefore, item parameters for all non-DIF items were constrained to possess identical item parameter values for each patient group.

All IRT models were fitted using an expectation-maximization algorithm with fixed quadrature (Chalmers, 2012). The Constrained Model and the Anchored Model were fitted using a Gaussian prior distribution of the latent trait, whereas the Final Model was fitted using the empirical histogram method for specifying the prior distribution of the latent trait as described by Bock & Aitkin (1981). To assess the assumption of unidimensionality, the following model fit indices were calculated and evaluated: Tucker-Lewis Index (TLI), Comparative Fit Index (CFI) and Root Mean Square Error of Approximation (RMSEA). The fit indices were evaluated against the cut-off values recommended by Hu & Bentler (1999).

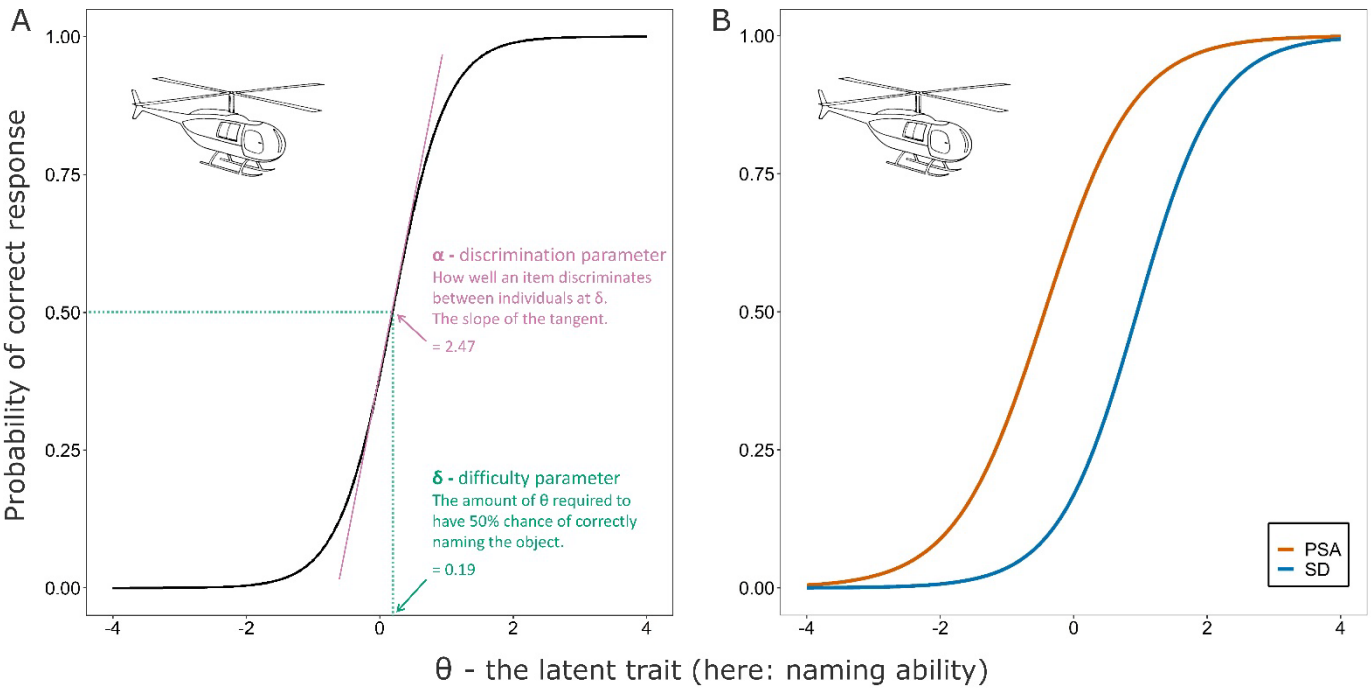

Figure 1. *Item characteristic curves (ICC) to illustrate the basics of IRT (A) and an example of DIF (B).* A: ICC of the item *helicopter*. The parameters visualized here are derived from the Constrained Model, in which item parameters are constrained to be equal across all patients. B: ICC of the item *helicopter* to illustrate DIF. The parameters visualized here are derived from the Final Model. For SD patients, the drawing of a helicopter requires more of the underlying trait to accurately name (δ = 0.95) than for PSA patients (δ = -0.44).

2.3.2 Regression analyses

In order to investigate which aspects of an item influence its difficulty and discrimination, a set of linear multiple regressions was computed for the two patient samples separately. The aim of these analyses was to determine which psycholinguistic and other variables were significantly related to the item parameters, and to what extent. Item difficulty or discrimination, respectively, were the dependent variables and the following independent variables were included: 1) word length, 2) frequency (HAL study, Lund & Burgess (1996)), 3) age of acquisition (Kuperman et al., 2012), 4) semantic diversity (Hoffman et al., 2013), 5) familiarity (Rossion & Pourtois, 2004), and 6) naming reaction time (RT) in healthy controls (Torrance et al., 2018). Variables 1-4 were obtained from the English Lexicon Project Web Site (Balota et al., 2007). Familiarity was taken from the normative data collection study by Rossion & Pourtois (2004), while naming RT in healthy controls was

obtained from the Multilanguage Written Picture Naming Dataset (MWPND; Torrance et al., 2018). Since some items, e.g. *lorry*, have several competing correct answers (*lorry*, *truck*), the respective values were extracted from the word that was most commonly given as an answer according to the MWPND (Torrance et al., 2018). One item consists of two words (*watering can*) and was not part of the English Lexicon Project, therefore the regressions are based on data for 63 items only.

Additionally, a follow-up regression analysis with Group (SD vs. PSA) and all Group x Item property interaction terms was conducted. The rationale for this analysis was that if an item property appeared to have substantially different importance for an item parameter between the two patient groups, then the difference needed to be formally assessed via the Group x Item property interaction. In the end, this follow-up analysis was only carried out for item difficulty as the regressions with item discrimination as dependent variable performed too poorly (see Results section for more details).

3. Results

3.1 Item analyses

Our first two aims were to investigate if a gradient of item naming difficulty could be established and whether this gradient varied between different aetiologies (PSA vs. SD). To this end, a series of IRT models was fitted to the individual response data of the two patient groups, resulting in a Final Model with two sets of item parameters (difficulty and discrimination) per group. The fit indices (RMSEA = .03, CFI = .99 and TLI = .99) of the Final Model indicated an excellent fit to a unidimensional structure when comparing these values to the recommended cut-offs by (Hu & Bentler, 1999). This suggests that the CSB naming test measures one underlying trait and qualifies the further analysis of individual item parameters.

Even though the items were widely distributed in terms of item difficulty for both patient group models, the gradient differed between the two clinical populations. These differences can be

inspected in the upper panel of Figure 2 – *helicopter*, for instance, contains a higher difficulty parameter in the SD ($\delta = 0.95$) than in the PSA ($\delta = -0.44$) sample. Similarly, items discriminated well across different levels of the underlying trait, but between-group differences in discrimination power were also present, as shown in the lower panel of Figure 2 (discrimination parameter, α, for *helicopter* was 1.68 and 1.49 for the SD and PSA sample, respectively).

The between-group disparities in item parameters can be formally investigated by looking at which items show DIF. Figure 1B shows the ICC curves for our example item (*helicopter*) which is one of the items that shows DIF. Thus, given the same level of naming impairment, *helicopter* is systematically easier to name for an individual with PSA versus an individual with SD. In total, DIF was significant in 27 items (10 in favour of PSA, 17 in favour of SD patients). This suggests that approximately half of the items' parameters differ across the two patient groups. More details on the individual item parameters can be found in the Supplementary Table S1.

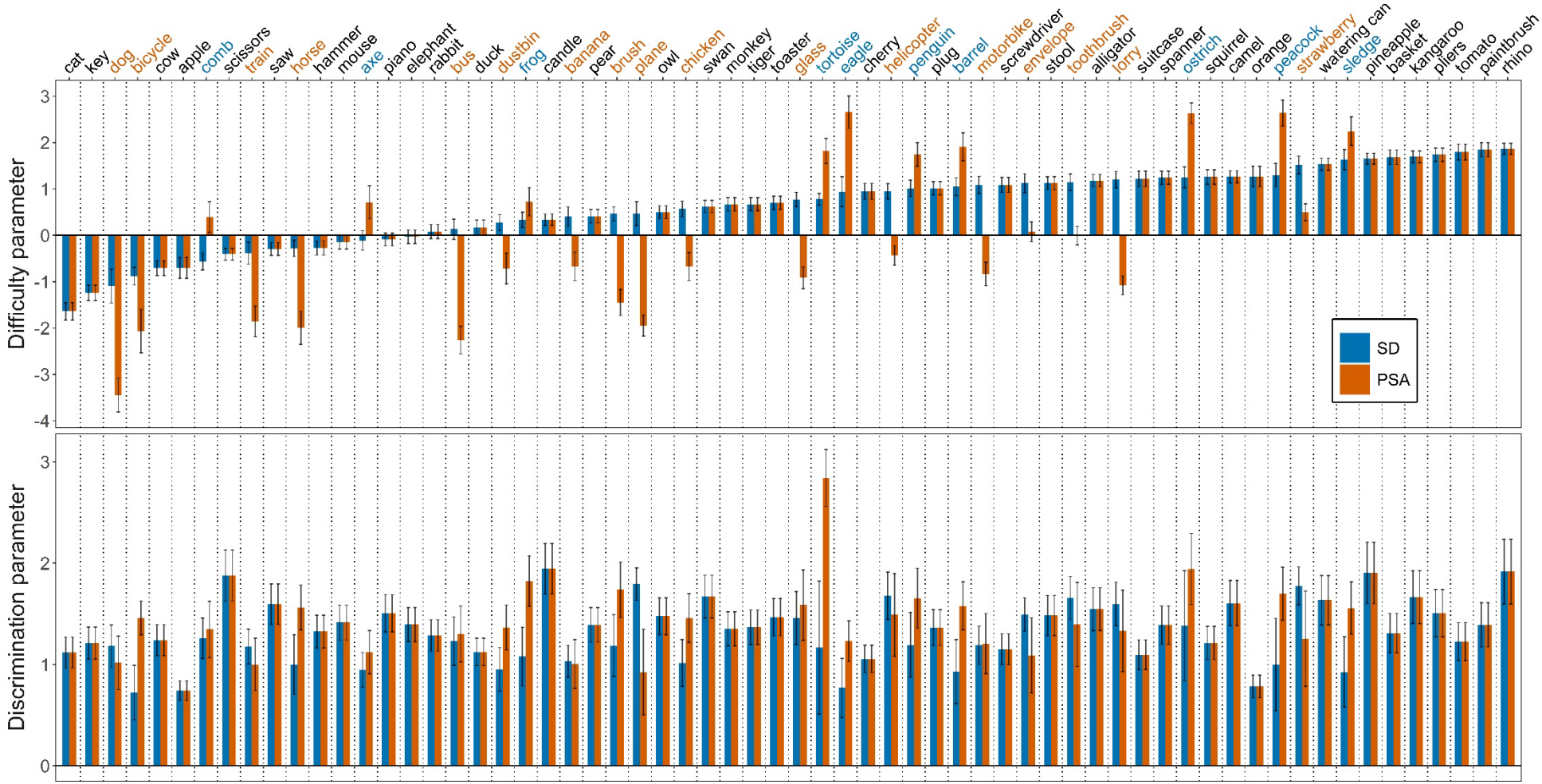

Figure 2. *Difficulty and discrimination parameters of each item for the Final IRT models.* Items are sorted by increasing difficulty based on the SD item parameters. Error bars indicate Standard Error of the estimated item parameter. Coloured words indicate that the DIF was significant. Blue-coloured words indicate that the item is systematically easier for SD than for PSA patients, orange-coloured words vice versa.

3.2 Regression analyses

Next, we aimed to characterize the extent to which psycholinguistic and other variables contribute to an item's naming difficulty and discrimination as well as to elucidate possible group differences in these patterns of relevant factors. Linear multiple regressions were computed for the two patient groups separately, including item difficulty or discrimination, respectively, as dependent variables. One behavioural (RT in healthy controls) and five psycholinguistic variables (word length, frequency, age of acquisition, semantic diversity, familiarity) were included as independent variables. The two regressions with item difficulty as dependent variable were significant (SD: $F(7,55) = 10.99$, $p < .001$; PSA: $F(6,56) = 12.52$, $p < .001$) and accounted for half of the variance (adjusted R^2 values of .54 and .53 for the SD and PSA, respectively). Inspection of the standardized regression coefficients revealed two significantly contributing variables for each analysis. Frequency was the most significantly contributing variable in both patient groups, while familiarity was also significant in the SD and word length in PSA. The absolute standardized regression coefficients are visualized in Figure 3. To test whether these variables contribute differentially across patient groups, a follow-up regression was computed by adding group as an independent variable and all Group x Item property interaction terms. The only significant interaction effect was Group x Familiarity ($t = -2.87$, $p = .005$), indicating that item familiarity plays a significantly larger role in item difficulty for SD patients compared to PSA patients.

The two regression analyses with item discrimination as dependent variable explained a negligible portion of variance (adjusted R^2 values of .13 and .05 for the SD and PSA, respectively), and the results of these analyses are thus not reported further in this paper.

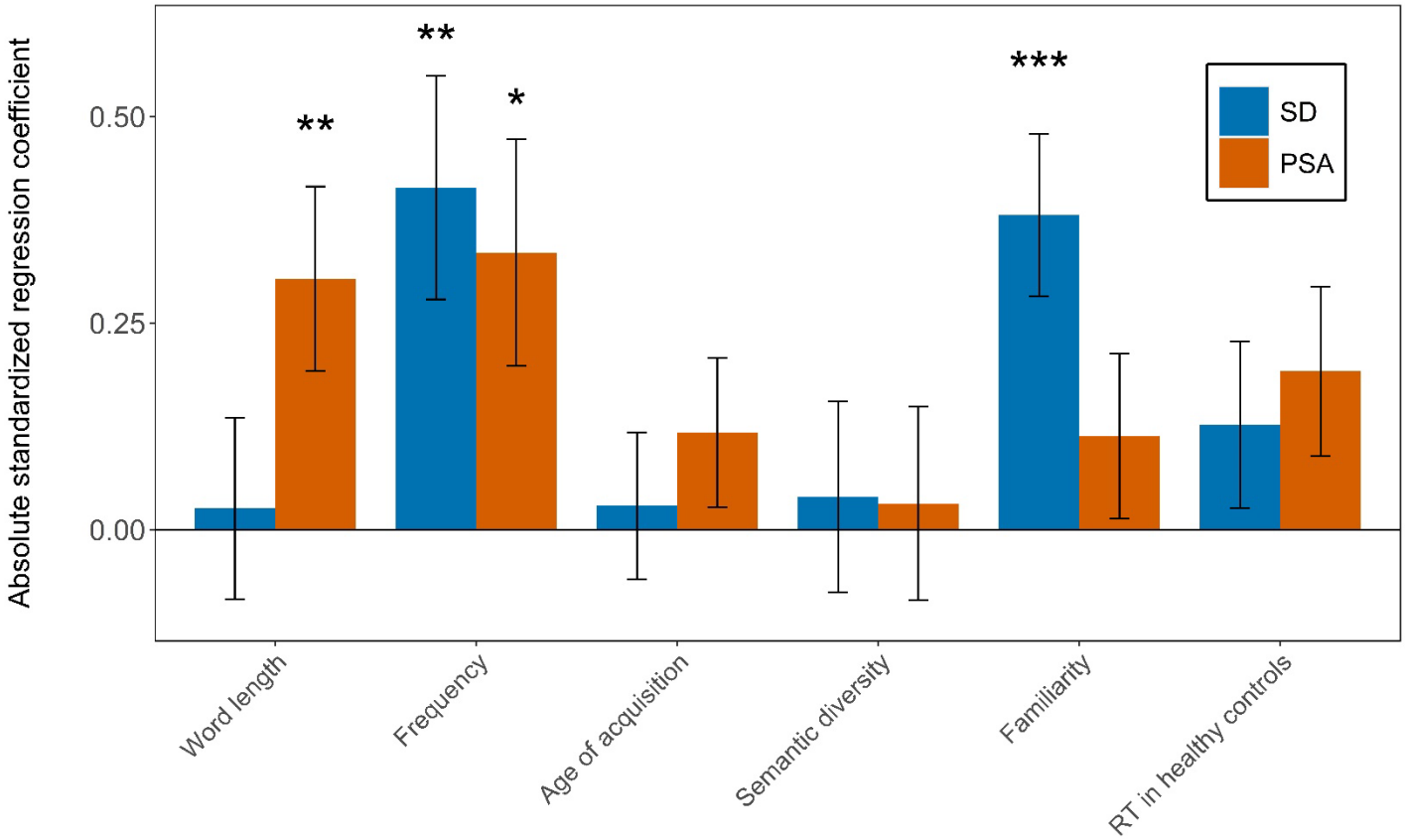

Figure 3. *Absolute standardized regression coefficients of the regression models on item difficulty for both patient groups.* Coefficients are absolute values. Asterisks (*, **, ***) indicate significant difference from zero at α = .05, .01 and .001 levels, respectively. Error bars indicate Standard Error of the Estimate.

4. Discussion

We investigated the existence and nature of a systematic gradient of item naming difficulty in semantic dementia and post-stroke aphasia, two neurological conditions with anomia as a central symptom. By employing Item Response Theory, we (a) established that an item naming difficulty gradient exists, which (b) partly differs between patient groups, and is (c) partially related to a limited number of psycholinguistic properties - frequency and familiarity for SD, frequency and word length for PSA.

4.1 A systematic gradient of item naming difficulty exists but differs across aetiology

Confirming clinical intuition, our results provide compelling evidence for the presence of a systematic gradient of item naming difficulty in both patient groups. The existence of a gradient means that the probability of successfully naming an item is not only dependent on overall anomia severity but also on the specific item to be named. If we imagine two patients with severe anomia who both name three out of ten pictures correctly, it is probable that they will get the same items correct or incorrect, respectively. The difficulty gradient was, however, not exactly the same between the two patient groups studied here. Approximately half of the items were significantly harder or easier to name for an SD patient versus a PSA patient with the same underlying naming ability.

4.2 A limited number of psycholinguistic item properties influences item difficulty

Having established that an item naming difficulty gradient exists but partly differs across patient groups, we then explored how strongly the gradient was related to various

psycholinguistic variables, and if these relationships varied across aetiologies. The regression analyses were significant for SD and PSA and explained over 50% of the variance. Frequency was the variable that was most strongly related to an item's difficulty in both patient groups, suggesting that words encountered more often are less susceptible to language impairments in general. In PSA patients, word length was the only other significant variable. These findings are largely in line with Fergadiotis et al. (2015), where word length, frequency plus additionally age of acquisition explained 62% of the variance of item naming difficulty in a PSA sample. The relative importance of word length for item difficulty in PSA patients in particular can be explained by (i) commonly occurring phonological impairment in most PSA patients (Halai et al., 2017) which is sensitive to word length (Crisp & Lambon Ralph, 2006; Nickels & Howard, 2004); and (ii) the co-occurring motor speech impairments in many PSA patients with anterior damage (Ziegler et al., 2022).

In SD patients the only additional contributor to item difficulty – over and above frequency – was familiarity. This result is in line with previous research which has found that SD patients' production in naming and connected speech, as well as comprehension are strongly related to a concept's familiarity (Bird et al., 2000; Lambon Ralph et al., 1998; Rogers et al., 2015). There may be at least two underpinning sources of this familiarity effect in SD. Premorbid conceptual representations are stronger when the concept is reinforced more often during learning and thus are more robust (but not immune) to the effects of ATL-centred atrophy (as observed in formal computational models of semantic memory and its decline in semantic dementia (Rogers et al., 2004)). Secondly, even during decline, the patients' ongoing experience may drive partial reinforcement, which will be greatest for the most commonly occurring,

familiar concepts. Accordingly, familiarity effects can become augmented during decline (Welbourne et al., 2011).

As for the approximately 50% of the variance in item difficulty that remain unexplained, several factors might be of relevance. First, successful picture naming relies on a multitude of elements, ranging from visually processing the stimulus, to activating the appropriate semantic content and accessing/selecting the correct word form, through to accurate articulation. The psycholinguistic properties included in our analyses only cover a limited number of these required elements, so others (for instance properties of the picture, influencing visual processing) are not considered. Furthermore, impairments in other non-language cognitive functions (such as attention or executive function) or other patient-related factors (such as age and education) might affect item-level performance, and thus influence the estimated item parameters. Finally, all of these elements, both language and non-language, would be subject to individual differences and thus add noise to the data.

4.3 Implications, limitations and future directions

Our findings have important implications relating to test interpretation, construction and beyond. First, the existence of a difficulty gradient offers an exciting opportunity to construct potentially more precise, tailored and efficient assessments. For instance, the gradients could be used for adaptive testing procedures whereby probe items of a certain difficulty level are chosen and the next item would depend on the success of naming the previous item. A similar approach could also be used in therapy settings, where the gradient would be helpful for choosing which pool of items would ideally be worked on next and thus constitute the “zone of proximal re-development” (Conroy et al., 2012).

Second, the finding of partly different difficulty gradients depending on aetiology generates both challenges and opportunities. Due to the differences between groups, it might be questionable as to whether it is sensible and justified to use the same assessment for different aetiology groups. At least when it comes to interpretation, the same total score might, at worst, not signify a comparable level of anomia severity (or naming ability) in one group versus another. Also, different gradients (i.e., a different item order) would have to be used for adaptive test construction. On the other hand, these group differences could potentially also be used to selectively construct tests that may help with differential diagnosis.

Third, knowledge about the factors that contribute most to an item's difficulty could be useful for estimating the difficulty of items that were not part of the studied item pool. This might be useful for the creation of parallel test versions or for an entirely new approach to tests that are constructed to systematically sample the variables in question.

The current investigation was limited to one naming test and two samples of patients. While the sample sizes are large for studies with such patient cohorts, they are on the smaller size for studies employing an IRT approach. As a consequence, our IRT results have higher margins of error compared to other IRT studies, e.g. Pedraza et al. (2011). Also, the PSA sample size was not large enough to explore potential differences across subgroups (e.g., comparing fluent and non-fluent patients).

Future studies should investigate to what extent the current findings hold in other (ideally larger) samples, with a different selection of items, a broader coverage of (psycholinguistic) item properties, and in languages other than English. In doing so it will be worth ensuring a broad coverage of anomia severity in the sample, as was true in the current study for both PSA and SD patients.

Author contributions

Erling Nørkær: Conceptualization, Methodology, Formal analysis, Visualization, Writing – Original Draft; Ajay D. Halai: Conceptualization, Methodology, Investigation, Writing – Review & Editing; Anna Woollams: Investigation, Writing – Review & Editing; Matthew A. Lambon Ralph: Conceptualization, Methodology, Supervision, Project administration, Writing – Original Draft; Rahel Schumacher: Conceptualization, Methodology, Supervision, Project administration, Writing – Original Draft.

Acknowledgements

We would like to thank the patients and their carers for contributing to this research project.

Funding

This research was supported by a grant from the Independent Research Fund Denmark (grant number DFF-024-00139B) to Randi Starrfelt, by an MRC Programme grant to MALR (MR/R023883/1), an MRC Career Development Award to ADH (MR/V031481/1), and an intramural award (MC_UU_00005/18).

Conflict of interests

The authors declare no conflicts of interest concerning the research, authorship, or publication of this study.

Open access

For the purpose of open access, the UKRI-funded authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

Adlam, A.-L. R., Patterson, K., Bozeat, S., & Hodges, J. R. (2010). The Cambridge Semantic Memory Test Battery: Detection of semantic deficits in semantic dementia and Alzheimer's disease. *Neurocase*, *16*(3), 193–207. https://doi.org/10.1080/13554790903405693

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Bird, H., Lambon Ralph, M. A., Patterson, K., & Hodges, J. R. (2000). The Rise and Fall of Frequency and Imageability: Noun and Verb Production in Semantic Dementia. *Brain and Language*, *73*(1), 17–49. https://doi.org/10.1006/brln.2000.2293

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Bozeat, S., Lambon Ralph, M. A., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Non-verbal semantic impairment in semantic dementia. *Neuropsychologia*, *38*(9), 1207–1215. https://doi.org/10.1016/S0028-3932(00)00034-8

Chalmers, R. P. (2012). **mirt**: A Multidimensional Item Response Theory Package for the *R* Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It Might Not Make a Big DIF: Improved Differential Test Functioning Statistics That Account for Sampling Variability. *Educational and Psychological Measurement*, *76*(1), 114–140. https://doi.org/10.1177/0013164415584576

Conroy, P. J., Snell, C., Sage, K. E., & Lambon Ralph, M. A. (2012). Using Phonemic Cueing of Spontaneous Naming to Predict Item Responsiveness to Therapy for Anomia in Aphasia. *Archives of Physical Medicine and Rehabilitation*, *93*(1), S53–S60. https://doi.org/10.1016/j.apmr.2011.07.205

Crisp, J., & Lambon Ralph, M. (2006). Unlocking the nature of the phonological-deep dyslexial continuum: The keys to reading aloud are in phonology and semantics. *Journal of Cognitive Neuroscience*, *18*. https://doi.org/10.1162/jocn.2006.18.3.348

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801–838. https://doi.org/10.1037/0033-295X.104.4.801

Ellis, A. W., Lum, C., & Lambon Ralph, M. A. (1996). On the use of regression techniques for the analysis of single case aphasic data. *Journal of Neurolinguistics*, *9*(3), 165–174. https://doi.org/10.1016/0911-6044(96)00008-5

Fergadiotis, G., Kellough, S., & Hula, W. D. (2015). Item Response Theory Modeling of the Philadelphia Naming Test. *Journal of Speech, Language, and Hearing Research*, *58*(3), 865–877. https://doi.org/10.1044/2015_JSLHR-L-14-0249

Goodglass, H., Kaplan, E., & Weintraub, S. (1983). *Boston naming test*. Lea & Febiger Philadelphia, PA.

Graves, R. E., Bezeau, S. C., Fogarty, J., & Blair, R. (2004). Boston Naming Test Short Forms: A Comparison of Previous Forms with New Item Response Theory Based Forms. *Journal of Clinical and Experimental Neuropsychology*, *26*(7), 891–902. https://doi.org/10.1080/13803390490510716

Halai, A. D., Woollams, A. M., & Lambon Ralph, M. A. (2017). Using principal component analysis to capture individual differences within a unified neuropsychological model of chronic post-stroke aphasia: Revealing the unique neural correlates of speech fluency, phonology and semantics. *Cortex*, *86*, 275–289. https://doi.org/10.1016/j.cortex.2016.04.016

Halai, A. D., Woollams, A. M., & Lambon Ralph, M. A. (2020). Investigating the effect of changing parameters when building prediction models for post-stroke aphasia. *Nature Human Behaviour*, *4*(7), 725-735. https://doi.org/10.1038/s41562-020-0854-5

Hodges, J. R., & Patterson, K. (2007). Semantic dementia: A unique clinicopathological syndrome. *The Lancet Neurology*, *6*(11), 1004–1014. https://doi.org/10.1016/S1474-4422(07)70266-1

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huber, W., Poeck, K., Weniger, D., & Willmes, K. (1983). *Aachener Aphasie Test (AAT): Handanweisung*. Verlag für Psychologie Hogrefe.

Kohn, S. E., & Goodglass, H. (1985). Picture-naming in aphasia. *Brain and Language*, *24*(2), 266–283. https://doi.org/10.1016/0093-934X(85)90135-X

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. https://doi.org/10.1177/0013164414529792

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Lambon Ralph, M. A., Graham, K. S., Ellis, A. W., & Hodges, J. R. (1998). Naming in semantic dementia—What matters? *Neuropsychologia*, *36*(8), 775–784. https://doi.org/10.1016/S0028-3932(97)00169-3

Lambon Ralph, M. A., McClelland, J. L., Patterson, K., Galton, C. J., & Hodges, J. R. (2001). No Right to Speak? The Relationship between Object Naming and Semantic Impairment:Neuropsychological Evidence and a Computational Model. *Journal of Cognitive Neuroscience*, *13*(3), 341–356. https://doi.org/10.1162/08989290151137395

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. https://doi.org/10.3758/BF03204766

Mckenna, P., & Warrington, E. K. (1983). *Graded naming test: Manual*. NFER-Nelson.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016–1031. https://doi.org/10.1037/a0027934

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165–178. https://doi.org/10.1037/h0027366

Nickels, L., & Howard, D. (1995). Aphasic naming: What matters? *Neuropsychologia*, *33*(10), 1281–1303. https://doi.org/10.1016/0028-3932(95)00102-9

Nickels, L., & Howard, D. (2004). Dissociating Effects of Number of Phonemes, Number of Syllables, and Syllabic Complexity on Word Production in Aphasia: It's the Number of Phonemes that Counts. *Cognitive Neuropsychology*, *21*(1), 57–78. https://doi.org/10.1080/02643290342000122

Pedraza, O., Sachs, B. C., Ferman, T. J., Rush, B. K., & Lucas, J. A. (2011). Difficulty and Discrimination Parameters of Boston Naming Test Items in a Consecutive Clinical Series. *Archives of Clinical Neuropsychology*, *26*(5), 434–444. https://doi.org/10.1093/arclin/acr042

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Paedagogiske Institut. https://books.google.dk/books?id=aB9qLgEACAAJ

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and Deterioration of Semantic Memory: A Neuropsychological and Computational Investigation. *Psychological Review*, *111*(1), 205–235. https://doi.org/10.1037/0033-295X.111.1.205

Rogers, T. T., Patterson, K., Jefferies, E., & Lambon Ralph, M. A. (2015). Disorders of representation and control in semantic cognition: Effects of familiarity, typicality, and specificity. *Neuropsychologia*, *76*, 220–239. https://doi.org/10.1016/j.neuropsychologia.2015.04.015

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's Object Pictorial Set: The Role of Surface Detail in Basic-Level Object Recognition. *Perception*, *33*(2), 217–236. https://doi.org/10.1068/p5117

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Teresi, J. A., Wang, C., Kleinman, M., Jones, R. N., & Weiss, D. J. (2021). Differential Item Functioning Analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS®) Measures: Methods, Challenges, Advances, and Future Directions. *Psychometrika*, *86*(3), 674–711. https://doi.org/10.1007/s11336-021-09775-0

Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. *Psychological Assessment*, *31*(12), 1442–1455. https://doi.org/10.1037/pas0000597

Torrance, M., Nottbusch, G., Alves, R. A., Arfé, B., Chanquoy, L., Chukharev-Hudilainen, E., Dimakos, I., Fidalgo, R., Hyönä, J., Jóhannesson, Ó. I., Madjarov, G., Pauly, D. N., Uppstad, P. H., van Waes, L., Vernon, M., & Wengelin, Å. (2018). Timed written picture naming in 14 European languages. *Behavior Research Methods*, *50*(2), 744–758. https://doi.org/10.3758/s13428-017-0902-x

Welbourne, S. R., Woollams, A. M., Crisp, J., & Lambon Ralph, M. A. (2011). The role of plasticity-related functional reorganization in the explanation of central dyslexias. *Cognitive Neuropsychology*, *28*(2), 65–108. https://doi.org/10.1080/02643294.2011.621937

Woollams, A. M., Cooper-Pye, E., Hodges, J. R., & Patterson, K. (2008). Anomia: A doubly typical signature of semantic dementia. *Neuropsychologia*, *46*(10), 2503–2514. https://doi.org/10.1016/j.neuropsychologia.2008.04.005

Ziegler, W., Aichert, I., Staiger, A., Willmes, K., Baumgaertner, A., Grewe, T., Flöel, A., Huber, W., Rocker, R., Korsukewitz, C., & Breitenstein, C. (2022). The prevalence of apraxia of speech in chronic aphasia after stroke: A bayesian hierarchical analysis. *Cortex*, *151*, 15–29. https://doi.org/10.1016/j.cortex.2022.02.012

Supplementary material

Table S1 in the supplementary material contains all the individual item parameters as well as indicators of how well each item fits the Final Model (item RMSEA and associated *p*-values). Additionally, the table contains information about which items contain DIF (based on the analyses of the differences in item parameters in the Anchored Model), as well as a measure of the effect size and direction of the DIF (Expected Score Standardized Difference (ESSD)). The ESSD is equivalent to a Cohen's *d* measure of the difference in expected test values between the two patient groups. In general, items fit adequately to the Final Model.

Table S1. Item parameters for the SD and PSA IRT model.

	PSA IRT Final Model				SD IRT Final Model				DIF analyses	
Item	Difficulty parameter	Discrimination parameter	Item fit – RMSEA	Item fit – *p* value	Difficulty parameter	Discrimination parameter	Item fit – RMSEA	Item fit – *p* value	DIF?	ESSD
helicopter	-0.44	1.49	0.00	0.55	0.95	1.68	0.11	0.07	DIF	-0.46
mouse	-0.15	1.41	0.00	0.68	-0.15	1.41	0.10	0.18	No DIF	0.05
toaster	0.70	1.47	0.00	0.83	0.70	1.47	0.00	0.52	No DIF	0.09
strawberry	0.50	1.25	0.03	0.35	1.52	1.78	0.07	0.23	DIF	-0.49
suitcase	1.22	1.10	0.00	0.46	1.22	1.10	0.10	0.07	No DIF	-0.38
cat	-1.64	1.12	0.05	0.11	-1.64	1.12	0.26	0.01	No DIF	-0.05
bicycle	-2.07	1.46	0.02	0.41	-0.88	0.72	0.06	0.25	DIF	-0.38
apple	-0.70	0.74	0.00	0.66	-0.70	0.74	0.08	0.15	No DIF	-0.09
rabbit	0.08	1.29	0.06	0.12	0.08	1.29	0.14	0.01	No DIF	-0.34
sledge	2.24	1.56	0.04	0.29	1.63	0.93	0.05	0.29	DIF	0.31
dustbin	-0.71	1.37	0.04	0.18	0.28	0.95	0.10	0.09	DIF	-0.35
frog	0.72	1.82	0.00	0.57	0.33	1.08	0.10	0.09	DIF	0.06

	PSA IRT Final Model				SD IRT Final Model				DIF analyses	
Item	Difficulty parameter	Discrimination parameter	Item fit – RMSEA	Item fit – *p* value	Difficulty parameter	Discrimination parameter	Item fit – RMSEA	Item fit – *p* value	DIF?	ESSD
tomato	1.80	1.23	0.01	0.43	1.80	1.23	0.07	0.19	No DIF	-0.18
lorry	-1.08	1.33	0.04	0.26	1.20	1.60	0.11	0.07	DIF	-0.76
cow	-0.71	1.24	0.00	0.87	-0.71	1.24	0.02	0.36	No DIF	0.06
watering can	1.53	1.63	0.04	0.26	1.53	1.63	0.14	0.02	No DIF	-0.25
pineapple	1.65	1.90	0.00	0.61	1.65	1.90	0.00	0.72	No DIF	0.00
bus	-2.26	1.30	0.00	0.44	0.13	1.23	0.05	0.30	DIF	-0.58
stool	1.13	1.48	0.05	0.18	1.13	1.48	0.09	0.14	No DIF	0.06
dog	-3.45	1.02	0.05	0.19	-1.10	1.18	0.07	0.25	DIF	-0.46
cherry	0.95	1.06	0.00	0.68	0.95	1.06	0.05	0.32	No DIF	-0.29
basket	1.69	1.31	0.00	0.67	1.69	1.31	0.10	0.10	DIF	0.55
train	-1.85	1.00	0.03	0.30	-0.39	1.18	0.00	0.47	DIF	-0.31
squirrel	1.26	1.21	0.06	0.10	1.26	1.21	0.04	0.35	No DIF	-0.18
pear	0.41	1.39	0.00	0.82	0.41	1.39	0.00	0.79	No DIF	0.02
horse	-2.00	1.56	0.00	0.44	-0.28	1.00	0.19	0.00	DIF	-0.44
motorbike	-0.84	1.21	0.00	0.66	1.09	1.19	0.02	0.41	DIF	-0.69
banana	-0.67	1.00	0.08	0.01	0.40	1.03	0.10	0.10	DIF	-0.34
barrel	1.91	1.58	0.00	0.67	1.05	0.93	0.04	0.34	DIF	0.37
plane	-1.95	0.92	0.00	0.91	0.46	1.79	0.16	0.04	DIF	-0.54
orange	1.27	0.78	0.02	0.38	1.27	0.78	0.10	0.05	No DIF	-0.17
piano	-0.09	1.50	0.03	0.35	-0.09	1.50	0.10	0.11	No DIF	-0.33
tortoise	1.82	2.84	0.00	0.46	0.78	1.17	0.00	0.87	DIF	0.42
pliers	1.74	1.51	0.00	0.47	1.74	1.51	0.08	0.19	No DIF	0.21
key	-1.25	1.21	0.03	0.31	-1.25	1.21	0.10	0.17	No DIF	0.05
penguin	1.74	1.65	0.00	0.70	1.01	1.19	0.06	0.24	DIF	0.29
axe	0.71	1.12	0.00	0.81	-0.11	0.95	0.05	0.29	No DIF	0.28

	PSA IRT Final Model				SD IRT Final Model				DIF analyses	
Item	Difficulty parameter	Discrimination parameter	Item fit – RMSEA	Item fit – *p* value	Difficulty parameter	Discrimination parameter	Item fit – RMSEA	Item fit – *p* value	DIF?	ESSD
monkey	0.67	1.35	0.00	0.76	0.67	1.35	0.07	0.21	No DIF	0.06
toothbrush	-0.01	1.40	0.00	0.59	1.14	1.66	0.06	0.25	DIF	-0.45
eagle	2.66	1.23	0.00	0.43	0.94	0.77	0.04	0.33	DIF	0.89
saw	-0.29	1.60	0.07	0.05	-0.29	1.60	0.08	0.22	No DIF	0.00
rhino	1.86	1.92	0.04	0.26	1.86	1.92	0.22	0.00	No DIF	0.35
plug	1.01	1.36	0.04	0.24	1.01	1.36	0.13	0.03	No DIF	-0.14
chicken	-0.68	1.46	0.00	0.63	0.57	1.02	0.05	0.30	DIF	-0.45
spanner	1.24	1.39	0.00	0.99	1.24	1.39	0.06	0.26	No DIF	-0.03
kangaroo	1.69	1.66	0.00	0.63	1.69	1.66	0.00	0.56	No DIF	0.15
glass	-0.91	1.59	0.03	0.29	0.77	1.46	0.06	0.26	DIF	-0.52
duck	0.16	1.13	0.00	0.85	0.16	1.13	0.19	0.00	No DIF	-0.10
scissors	-0.41	1.88	0.03	0.33	-0.41	1.88	NA	NA	No DIF	0.00
camel	1.26	1.61	0.00	0.96	1.26	1.61	0.30	0.00	No DIF	0.56
envelope	0.07	1.09	0.00	0.80	1.13	1.49	0.02	0.41	DIF	-0.41
owl	0.50	1.48	0.06	0.10	0.50	1.48	0.18	0.01	No DIF	0.21
paintbrush	1.85	1.39	0.00	0.46	1.85	1.39	0.11	0.09	No DIF	-0.66
tiger	0.67	1.37	0.06	0.14	0.67	1.37	0.00	0.57	No DIF	-0.13
comb	0.39	1.35	0.07	0.06	-0.57	1.26	0.12	0.13	DIF	0.30
swan	0.62	1.67	0.00	0.92	0.62	1.67	0.15	0.04	No DIF	0.00
screwdriver	1.09	1.15	0.01	0.42	1.09	1.15	0.00	0.42	No DIF	-0.51
elephant	-0.03	1.39	0.00	0.85	-0.03	1.39	0.00	0.63	No DIF	-0.24
candle	0.34	1.94	0.00	0.47	0.34	1.94	0.00	0.47	No DIF	0.00
ostrich	2.63	1.94	0.05	0.22	1.25	1.38	0.00	0.45	DIF	0.76
alligator	1.18	1.55	0.00	0.73	1.18	1.55	0.00	0.91	No DIF	-0.01
brush	-1.45	1.74	0.00	0.85	0.46	1.19	0.08	0.19	DIF	-0.55

	PSA IRT Final Model				SD IRT Final Model				DIF analyses	
Item	Difficulty parameter	Discrimination parameter	Item fit – RMSEA	Item fit – *p* value	Difficulty parameter	Discrimination parameter	Item fit – RMSEA	Item fit – *p* value	DIF?	ESSD
peacock	2.64	1.70	0.00	0.82	1.30	1.00	0.02	0.42	DIF	0.73
hammer	-0.27	1.33	0.06	0.10	-0.27	1.33	0.11	0.14	No DIF	-0.12

Note: The items *pineapple*, s*aw*, s*cissors*, s*wan* and *candle* were used as anchor items in the IRT model.

Item fit RMSEA: Root Mean Square Error of Approximation. Item fit *p* values are associated with χ^2 tests testing for significant lack of fit.

DIF: Differential Item Functioning.

ESSD: Expected Score Standardized Difference (equivalent to Cohen's *d* for the difference in expected test scores between patient groups). Negative values indicate that an item is systematically easier for PSA patients than for SD patients.