Reading tasks from PALPA:

How do controls perform on Visual Lexical decision, Homophony, Rhyme, and Synonym Judgements?

Lyndsey Nickels

&

Jennifer Cole-Virtue

Macquarie Centre for Cognitive Science (MACCS)

Macquarie University, Sydney

Australia.

Aphasiology, in press.

Running Head: PALPA reading tasks
Keywords: PALPA, control, homophone judgements, rhyme judgements, visual lexical decision, synonym judgements

Address Correspondence to:
Lyndsey Nickels,
Macquarie Centre for Cognitive Science (MACCS),
Macquarie University, Sydney,
NSW 2109, Australia.
Tel: +61-2-9850-8448
Fax: +61-2-9850-6059
Email: lnickels@maccs.mq.edu.au

ABSTRACT

Background: PALPA (Psycholinguistic Assessments of Language Processing in Aphasia; Kay, Lesser & Coltheart, 1992) is a resource widely used by both clinicians and researchers. However, several of the subtests lack data regarding the performance of proficient English language speakers on these tasks.

Aims: This paper investigates factors affecting the speed and accuracy of performance of young control participants on four assessments from PALPA: Visual lexical decision (subtest 25); synonym judgements (subtest 50); rhyme judgements (subtest 15); and homophone judgements (subtest 28).

Methods and Procedures: Data is presented regarding both speed and accuracy of performance on each of the four tasks, and statistical analysis of those factors which influence performance within each test is carried out, for the participants as a group and also for the individuals within the group.

Outcomes and Results: Visual Lexical decision showed significant effects of frequency on response latency and accuracy and of lexicality and imageability on response latency alone; synonym judgements showed significant effects of imageability on response latency; significant effects of word type were found on response latency for Homophone Judgements; for rhyme judgements there was a significant effect of rhyme for both accuracy and latency, and a significant interaction between rhyme and visual similarity.

Conclusions: For the clinician seeking to interpret the performance of the person with aphasia on the tasks we have described here, we have presented data which provides some indication of the speed and accuracy of performance of young controls on these tasks. It is clear that ceiling effects in accuracy mask effects of psycholinguistic variables on normal performance that become apparent when speed of response is considered. However, performance is far from at ceiling for all the tasks described - some participants perform close to chance on some conditions. Finally, these data highlight the fact that comparison of the

pattern of performance of individual participants with that of a group of controls can be problematic given the variability of control patterns of performance.

INTRODUCTION

PALPA (Psycholinguistic Assessments of Language Processing in Aphasia; Kay, Lesser & Coltheart, 1992) is a resource widely used by both clinicians and researchers. It has proved invaluable but is not without its weaknesses. One of these is the relative lack of data regarding how proficient English language users (so called “normal” speakers or “controls”) perform on many of the tasks (Basso, 1996; Marshall, 1996, Wertz, 1996). The authors suggest that one can assume that controls will perform at ceiling on the majority of the tasks – and indeed this may be true for tasks such as repetition and reading aloud. However, one is less confident regarding control performance on some of the more complex tasks (e.g. silent rhyme judgements) and those using abstract and low frequency vocabulary (e.g. synonym judgements). Hence, at least for these tasks control data is required.

Moreover, there is an argument that tasks where most controls are performing at ceiling may not be optimal in terms of evaluating the performance of the person with aphasia (Kay, Lesser & Coltheart, 1996). Best (2000) argues that performance with accuracy at ceiling may mask the fact that some subsets of stimuli in a task may be harder than others (e.g. visually similar rhyming pairs versus non visually similar rhyming pairs). This can make interpreting the performance of an individual with aphasia difficult. Furthermore, can we be confident that, by scoring at ceiling in terms of accuracy, the aphasic person is performing as they would have premorbidly? Is the presence of, for example, worse performance on visually similar rhyming pairs significant or is it merely that these are also more difficult for controls (even if performance is accurate)? One way of avoiding ceiling effects on performance is to examine both accuracy and speed of response on a task. Measuring speed of response allows the relative difficulty of subsets within tasks to be assessed even when accuracy is at ceiling, and gives a more sensitive measure of “normal” performance. This is particularly valuable when assessing those aphasic individuals with more subtle impairments.

Thus, in this paper we investigate the performance of young control participants on four assessments from PALPA: Visual lexical decision (subtest 25); synonym judgements (subtest 50); rhyme judgements (subtest 15); and homophone judgements (subtest 28). The latter

three assessments have no normative data provided, and visual lexical decision has accuracy data alone from 26 elderly participants (spouses of people with aphasia), and with some omissions regarding the overall performance (e.g. no overall measure of accuracy). Here we will present data regarding both speed and accuracy of performance on each of the four tasks, and statistical analysis regarding those factors influencing performance within each test. We will discuss the pattern shown by the groups of participants and also by individuals within the groups.

METHOD

Participants

The twenty-one participants in this study were all undergraduate students from Macquarie University who were speakers of Australian English. Seventeen were female and four were male, the average age was 25.4 years (age range of 19-48 years). The students participated in the experiment as part of the fulfilment of their course requirements, or for payment of AUD$10.

Materials

Four tasks were presented: visual lexical decision, synonym judgements, homophone judgements and rhyme judgements. Stimuli were taken from PALPA (Kay, Lesser & Coltheart, 1992; subtests 25[1], 50, 28 and 15, respectively).

Visual Lexical Decision Task (Subtest 25)

The aim of this task is for the participant to decide whether a written letter string is a word. The lexical decision task contained ten practice items and 120 test stimuli. The test items consisted of 60 nonwords and 60 words. The word stimuli were in four subsets of fifteen items each systematically varying imageability and frequency within the subsets (High imageability-High frequency, High imageability-Low frequency, Low imageability-High frequency, Low imageability-Low frequency). Words are matched across groups (pairwise) as far as possible for grammatical class, number of letters, syllables and morphemes. Nonwords are derived from words by changing one or more letters, while preserving orthotactic and phonotactic

legality. The manipulation of frequency and imageability across sets allows the effects of these variables on performance to be evaluated.

Synonym Judgement Task (Subtest 50)

In this task the participant has to judge whether two written words are similar in meaning – approximately synonymous. There were four practice items (car-automobile, tree-house, help-code, start-beginning) and 60 test word pairs. Thirty of the stimulus items are words of high imageability and half of low imageability. Within each set, thirty of the items are (approximately) synonymous requiring a yes response and thirty are unrelated in meaning requiring a no response. The high and low imageability sets are matched for word frequency. The difference in imageability between the sets allows the effect of this variable on performance to be evaluated.

Homophone Judgement Task (Subtest 28)

In this task, the participant has to judge whether a written word pair (eg. prey-pray; bore-bow) or non-word pair (eg. heem-heam; bick-blic) sound the same. The homophone judgement task has four practice items and 60 test word pairs. There are three subsets with 20 stimulus pairs each; regular, exception, and non-word. Each subset comprises 10 homophonic and 10 non-homophonic word pairs. The non-homophonic pairs are matched for visual similarity to the homophonic pairs. This task allows the effect of stimulus type, lexicality and word regularity on the generation of phonology from print to be evaluated.

Rhyme Judgement Task (Subtest 15)

The aim of this task is for the participant to judge if two written words rhyme. To complete this task correctly the participant has to derive phonology from the written word, segment off the rime and compare the segmented stimuli. There were four practice items and 60 test word pairs, in four subsets of 15 words each. Half of the stimulus pairs rhyme and half are non-rhyming pairs. In the rhyming pairs, half the words (spelling pattern rhyme: SPR) share the same orthographic body and a decision based on visual similarity will result in a correct response (eg. town-gown). The other half (phonological rhyme: PR), comprise rhyming pairs that have different orthographic bodies and in these cases a correct judgement can only be

made if the participant knows how the word sounds (eg. bowl-mole). The non-rhyming pairs are also in two halves, half share the same orthographic bodies (spelling pattern control: SPC) and here the visual similarity may mislead (eg. down-flown). The remaining half of the non-rhyming pairs (phonological control: PC) are visually dissimilar, and also share the same bodies as the rhyming pairs (eg. hoe-chew, corresponding with shoe-screw in the rhyming pairs).

Hence these subsets allow the effect of visual similarity between the word pairs to be assessed in the rhyme and non-rhyme conditions. Here we have chosen to use more descriptive (and hopefully transparent) labels for these PALPA subsets, reflecting the rhyme and visual similarity manipulation: SPR - rhyme-vissim (rhyme, visually similar); PR – rhyme-novissim (rhyme, not visually similar); SPC – norhyme-vissim (no rhyme, visually similar); PC-norhyme-novissim (no rhyme, not visually similar).

Apparatus

The experimental control programme DMDX (Forster & Forster,2003) running on a Pentium III PC was used for presentation of the stimuli and the recording of responses for all four tasks.

Procedure

Participants were tested individually and required to sit approximately 14 inches away from the computer monitor. All 4 tasks were presented in a single session with order of task presentation randomised across participants. However, due to individual testing constraints not all participants completed all four tasks (and equipment error resulted in some participant data being lost). Task instructions were given verbally by the tester and also visually on the computer screen.

For example, the instructions for homophone judgement were as follows;

"For this task, you will see pairs of words or nonwords,

your job is to decide if they sound the same,

as quickly as you can, without making errors,

DO NOT SAY THE WORDS ALOUD

If they sound the same, press +,

If they DO NOT sound the same, press -,

Press NEXT to start practice".

Instructions were essentially of the same format for all tasks with only the first two lines changing for each task; lexical decision "you will see a letter string, your job is to decide if the letter string is a real word or a non word", rhyme judgments "you will see pairs of words, your job is to decide if the words rhyme", and for synonym judgements "decide if the words are similar in meaning".

Participants were instructed to make their decision as quickly as they could and press a + or – button on a response pad to indicate their decision. To indicate a yes response participants had to respond by pressing the + button and the - button for a no response. Each task had a number of practice items and the tester provided feedback following completion of these items. The number of practice and stimulus items varied across tasks, as noted above.
The participant was then instructed to continue to the test items. The inter stimulus interval for all tasks was one second.

RESULTS

Visual Lexical Decision Task

Group analyses

Reaction time and error data are presented in table 1 (details of errors per item can be found in Appendices B and C). These data were analysed by-subjects and by-items using analysis of variance (ANOVA). In the by-subjects analysis the factors of lexicality (word, non-word), imageability (high imageability, low imageability) and frequency (high frequency, low frequency) were treated as repeated measures and used to evaluate mean reaction time and accuracy per participant. In the by-items analysis these same factors were treated as independent measures when used to evaluate mean reaction time and accuracy per item.

*****Table 1 about here*****

Reaction Time

There was a significant effect on mean reaction time of lexicality by-subjects (F 1,20)=61.75, p=0.000) and by-items (F(1,118)=88.33, p=0.000). Participants were faster to respond to words (requiring a yes response) than non-words (requiring a no response). There was no significant effect of lexicality on mean error either by-subjects (F(1,20)=1.38, p=0.255) or by-items (F(1,118)=0.76, p=0.386).

Within the yes responses (words) there was a significant effect on mean reaction time of imageability and frequency both by-subjects (imageability: F(1,20)=11.52, p=0.003; frequency : F(1,20)=23.86, p=0.000) and by-items (imageability: F(1,56)=4.73, p=0.034; frequency: F(1,56)=16.15, p=0.000). However, there was no significant interaction between imageability and frequency either by-subjects (F(1,20)=0.62, p=.0442) or by-items (F(1,56)=0.55, p=0.461).

Accuracy

There was no significant effect on accuracy of imageability by-subjects (F(1,20)=1.18, p=0.289) or by-items (F(1,56)=1.40, p=0.241). Frequency had a significant effect on accuracy both by-subjects (F(1,20)=13.61, p=0.001) and by-items (F(1,56)=16.65, p=0.000). There was no significant interaction between imageability and frequency on accuracy by-subjects (F(1,20)=2.25, p=0.149) or by-items (F(1,56)=2.10, p=0.153).

Individual analyses

Reaction Time

All but two of the participants (19/21, 90%) showed faster reaction times with high frequency than low frequency stimuli, and eight participants (38%) showed a significant advantage for high frequency stimuli. No participant showed a significant advantage for low frequency stimuli and those who showed numerically faster mean reaction times showed very small differences (6 ms and 14 ms).

More participants showed faster responses to high imageability stimuli than to low imageability stimuli (18/21, 86%) but few showed significant effects of imageability on performance (3/21, 14%). (Individual data can be found in Appendix D).

Accuracy

Participant performance was generally too close to ceiling to make statistical analysis of errors viable for most individuals. However, while only one individual participant showed a significant effect of frequency on accuracy, every participant who showed a difference between high and low frequency stimuli showed worse performance on low frequency stimuli, with only one exception (and this participant only made 2 errors). In contrast, while once again one participant showed a significant effect of imageability on accuracy, there was much more variability with 5 participants making more errors with low imageability than high imageability stimuli (as would be expected from the group analysis).

Synonym Judgement Task

Group analyses

One participant's data was excluded due to equipment failure.

Reaction time and error data are presented in table 2 (for details of errors per item see Appendix E). These data were analysed by-subjects and by-items using analysis of variance (ANOVA). In the by-subjects analysis the factors of imageability (high, low) and synonymy (synonymous, non-synonymous) were treated as repeated measures and used to evaluate mean reaction time and accuracy per participant. In the by-items analysis the factors of imageability (high, low) and synonymy (synonymous, non-synonymous) were treated as independent measures and used to evaluate mean reaction time and accuracy per item.

******** Table 2 about here ********

Reaction Time

There was a significant effect of imageability by-subjects (F(1,19)=65.91, p=0.000) and by-items (F(1,56)=38.25, p=0.000) on reaction time. Participants responded faster to high imageability than to low imageability items.

There was no effect of synonymy on reaction time by-subjects (F(1,19)==2.40, p=0.138) or by-items (F(1,56)=1.15, p=0.287). There was a significant interaction between synonymy and imageability by-subjects only (F(1,19)=6.29, p=0.021; by-items: F(1,56)=3.54, p=0.065). This interaction reflects the fact that for high imageability items responses to non-synonymous pairs was slower, and for low imageability items responses to synonymous pairs were slower.

Accuracy

There was a significant effect of imageability on accuracy by-subjects only (F (1,19)=21.40, p=0.000; By-items: F(1,56)=1.84, p=0.180) with higher accuracy on high imageability items. There was a significant effect of synonymy by-subjects (F (1,19)=11.43, p=0.003) and by-items (F(1,56)=4.54, p=0.037), with responses to non-synonymous pairs being more accurate. There was no interaction between imageability and synonymy by-subjects or items.

Individual Analyses

Reaction Time

Every individual participant within the group was faster to respond to high imageability than low imageability stimuli, and this was significant for the majority of the participants (70%; See appendix F). The mean effect size (Low imageability RT minus High Imageability RT) was 200.7msecs with 95% confidence limits from 152.7msecs to 248.7 msecs.

Accuracy

As error rates were low, statistical analysis was not performed. Only four individuals (20%) showed worse performance on low imageability stimuli and in all cases the difference was only one item.

Homophone Judgement Task

Group analyses

Reaction time and accuracy data are presented in table 3 (for details of errors per item see Appendix G). These data were analysed by-subjects and by-items using analysis of variance (ANOVA). Word type was further examined using related t-tests for by-subjects and independent t-tests for by-items analysis. In the by-subjects analysis the factors of word type (regular, exception/irregular and nonword) and homophony (homophonic, non-homophonic) were treated as repeated measures and used to evaluate mean reaction time and accuracy per participant. In the by-items analysis the factors of word type (regular, exception/irregular and nonword) and homophony (homophonic, non-homophonic) were treated as independent measures and used to evaluate mean reaction time and accuracy per item.

******Table 3 about here******

Reaction Time

There was a significant effect on mean reaction time of word type by-subjects ($F(2,40)=8.28$, $p=0.001$) and by-items ($F(2,54)=22.70$, $p=0.000$). There was also a significant effect of homophones on mean reaction time by-subjects ($F(1,20)=27.58$, $p=0.000$) and by-items ($F(1,54)=12.31$, $p=0.001$). Participants were faster to respond to items that required a yes response i.e. homophonic word pairs than non-homophonic pairs. There was a significant interaction between homophony and word type only by-subjects ($F(2,40)=50.14$, $p=0.000$); by-items ($F(2,54)=1.88$, $p=0.162$), reflecting the fact that regular words show a larger effect of homophony on reaction time than either irregular words or nonwords.

*****Table 4 about here******

Accuracy

There was no significant effect on accuracy by-subjects or by-items of word type (by-subjects: $F(2,40)=0.75$, $p=0.479$, by-items: $F(2,54)=0.53$, $p=0.593$) or homophones (by-subjects: $F(1,20)=1.22$, $p=0.283$, by-items: $F(1,54)=0.70$, $p=0.406$). There was also no significant interaction either by-subjects ($F(2,40)=2.84$, $p=0.070$) or by-items ($F(2,54)=0.29$, $p=0.749$) on accuracy.

The effect of word type was further analysed using paired (by-subjects) and independent (by-items) t-tests (see table 4). There were no significant differences between groups in accuracy, but in reaction time, regular and irregular word pairs were significantly faster than nonword pairs, both by subjects and by items. Regular pairs were significantly faster than irregular pairs by subjects but not by items.

Individual analyses

Reaction Time

71% of participants showed the effect of word type that was true of the group (regular word pairs faster than exception word pairs which are faster than nonword pairs; see Appendix H). All participants showed faster reaction times to regular words than to nonwords and this was significant for 71% of participants. Most participants responded faster to exception words than nonwords (only one did not and this was a very small difference – 12msecs), but this was only significant for nine participants (43%). 76% of participants showed faster responses to regular words but these effects were only significant for two individuals (10%).

Accuracy

As the group showed no significant effects on accuracy, individual analyses were not attempted.

Rhyme Judgement Task

Group analyses

Only seventeen participants performed this task. Reaction time and error data are presented in table 5 (for accuracy for each item see Appendix I). These data were analysed by-subjects and by-items using analysis of variance (ANOVA). In the by-subjects analysis the factors of rhyme (rhyme, non-rhyme) and visual similarity (visually similar, non-visually similar) were treated as repeated measures and used to evaluate mean reaction time and accuracy per participant. In the by-items analysis the factors of rhyme (rhyme, non-rhyme) and visual similarity (visually similar, non-visually similar) were treated as independent measures and used to evaluate mean reaction time and accuracy per item.

******Table 5 about here******

Reaction Time

There was a significant effect on mean reaction time of rhyme by-subjects ($F(16,1)=46.29$, $p=0.000$) and by-items ($F(1,56)=45.54$, $p=0.000$). Participants were faster to judge rhyming pairs than non-rhyming pairs. There was no significant effect on mean reaction time of visual similarity either by-subjects ($F(16,1)=.69$, $p=0.418$) or by-items ($F(1,56)=0.01$, $p=0.935$). There was a significant interaction between rhyme and visual similarity both by-subjects ($F(16,1)=11.29$, $p=0.004$) and by-items ($F(1,56)=5.88$, $p=0.019$).

Accuracy

There was a significant effect on accuracy of rhyme both by-subjects ($F(16,1)=15.50$, $p=0.001$) and by-items ($F(1,56)=18.12$, $p=0.000$). There was no significant effect of visual similarity on accuracy either by-subjects ($F(16,1)=1.92$, $p=0.185$) or by-items ($F(1,56)=0.72$, $p=0.398$). The interaction between rhyme and visual similarity for accuracy was significant by-subjects ($F(16,1)=13.39$, $p=0.002$) and by-items ($F(1,56)=7.210$, $p=0.010$).

********** Figure 1 about here *********

The interactions between rhyme and visual similarity are illustrated in Figure 1. For both reaction time and error, they reflect the fact that for rhyming items error rate and response time are both smaller when the stimuli are visually similar, in contrast for nonrhyming items error rate and response time are smaller when the pairs are visually dissimilar.

Overall, the fastest and most accurate pairs were visually similar rhymes, then visually dissimilar rhymes, then visually dissimilar nonrhymes, with visually similar non-rhymes being the slowest and most error prone (t-test results presented in Appendix A).

Individual Analyses

Reaction Time

Every participant showed numerically faster responses to rhymes compared to nonrhymes, and ten of the participants (59%) showed significant effects of rhyme using ANOVA (see

Appendix J). Again consistent with the group results no individual showed significant effects of visual similarity on reaction time. However, five individuals showed a significant interaction between rhyme and visual similarity, and eleven individuals showed the same pattern as the group with faster reaction times for visually similar rhymes and slower reaction times for visually similar nonrhymes.

Accuracy

Error rates were relatively high for some participants on this task. Indeed on some subsets some participants did not perform better than chance[2]. All participants scored above chance on rhyming pairs (both visually similar and non-visually similar). However, three participants scored no better than chance on non-rhyming pairs overall - with four performing no better than chance on non-visually similar non-rhymes and seven performing no better than chance on visually similar non-rhymes.

Five individuals showed significant effects of rhyme on accuracy, and all but one participant showed better performance with rhyming than non-rhyming pairs. No participant showed a significant effect of visual similarity on accuracy and no clear pattern emerged (as predicted from the group data).

DISCUSSION

********* Tables 6 & 7 about here *********

We have investigated the performance of young control participants on four tasks from PALPA. A summary of overall mean accuracy and reaction time for each task is presented in Table 6, and those factors that significantly affected young control participant performance are summarised in table 7. We will first summarise the results for each subtest before embarking on further discussion.

Visual Lexical decision

Participants were generally accurate on this task. There were no significant effects of lexicality or imageability on accuracy, although there was a significant effect of frequency with the group performing less accurately with low frequency stimuli. In contrast, there were significant effects not only of frequency, but also of imageability and lexicality on speed of response (Nickels & Cole-Virtue, 2004). Individuals generally showed the same pattern as the group, and no participant showed a significant (nor substantial) effect of frequency in the opposite direction to the group.

Synonym Judgements

While generally accurate, only two participants produced no errors on this task. Effects of imageability were found for reaction times but not for errors, and these were robust across individuals with no participant having slower reaction times to high imageability than low imageability stimuli.

Homophone Judgements

There was more variability of accuracy on this task, with some participants scoring relatively poorly, particularly on nonword pairs. There were no significant effects on accuracy but word type (regular, exception or nonword) significantly affected reaction time. The significantly faster response to regular words than to nonwords was robust across individuals.

Rhyme Judgements

This task showed the greatest variability in accuracy and some subjects showed performance at chance on some subsets. There was a significant effect of rhyme for both accuracy and latency, which was moderately consistent across individuals. There was a significant interaction between visual similarity and rhyme such that the group was faster and less error prone for visually similar rhyming pairs and slowest and most error prone for visually similar non-rhyming pairs (although this pattern was not clear for individuals). Hence orthography has a marked effect on this phonological judgement.

Comparisons between 'normal' and 'aphasic' performance

For the clinician seeking to interpret the performance of the person with aphasia on the tasks we have described here, we have presented data that provides some indication of the speed and accuracy of performance of young controls on these tasks. However, as Kay et al (1996) note these data cannot necessarily provide the answer to whether a particular individual with aphasia is performing on these assessments as they were premorbidly – this would require a group matched to that individual on, for example, age, educational history, occupation and cultural background. Nevertheless, these data do help us on our way to deciding "how many errors constitutes a deficit" (Marshall, 1996).

However, there are also some cautionary messages to take away from our investigations, not least that controls can perform surprisingly poorly on what are intuitively straightforward tasks.

Effects of variables on performance & inferring level of impairment

The discovery that a psycholinguistic variable affects the performance of the person with aphasia has frequently been interpreted to indicate an impairment of the stage of processing at which that variable is thought to operate. For example, an effect of frequency has been interpreted as evidence for a lexical level impairment, an effect of imageability for a semantic impairment. While it has been acknowledged that some of these variables also affect 'normal' (speed of) processing there has been little discussion of the implications of this fact. If, as is the case here, 'normal' subjects show effects of frequency on lexical decision (both for speed and accuracy) and imageability on synonym judgements, can we necessarily infer that the aphasic individual who shows an effect of frequency is necessarily impaired in lexical access? Might it not be the case that this individual is showing the same effect of frequency that is the consequence of the normal system (but perhaps with a reduced overall level of accuracy)?

Effects of variables on performance & individual variability.

In the experimental investigation of language processing with so-called 'normals', the standard methodology is to report group statistics, with little attention to the performance of individuals within the group. This is on the premise that the underlying language system is identical in humans (without language impairment and who are speakers of the same

language) but that data is inherently noisy. Hence, by averaging across a group of individuals the ‘noise’ is reduced and the ‘true’ picture emerges. The difficulty with this approach is that in the clinical setting one is faced with a single individual ‘noise and all’! One approach that is used in research is to reduce the noise by multiple assessments or using very large samples of behaviour; clinically this approach is impractical. Hence, here we presented data from the individuals within the group in an attempt to ascertain how robust the effects were across individuals. For most effects the answer is ‘not very’. The best that can be said is that no individual showed a significant result that was in the reverse direction to that of the group. Hence, little can be concluded from the lack of a significant (or absolute direction of) effect of variables on performance, but if an individual shows a significant effect in the reverse direction to that of the group results reported here that is more likely to be an indication of impairment.

Effect sizes and their relationship to overall speed of processing

********** Table 8 & figure 2 about here **********

We have already discussed the extent to which effects were reliable across individual participants, and the problem of interpreting the behaviour of a particular aphasic individual in relation to this (lack of) reliability. In table 8 we present another means of summarising the data – in terms of mean effect sizes and the 95% confidence limits for that mean. Hence, for example, for synonym judgements low imageability stimuli were responded to on average slower than high imageability stimuli. The mean difference between the reaction times, the effect size, was 201msecs. The upper confidence limit is 249msecs and the lower 153msecs. In other words, based on this sample, 95% of the population will be between 153 and 249 msecs slower to respond to low imageability stimuli in this synonym judgements task than to high imageability stimuli. It might, therefore seem reasonable to conclude that an individual with aphasia who shows an effect size outside these limits is not performing ‘normally’. Unfortunately this may be overly simplistic. Figure 2 shows the relationship between overall speed of response (overall mean RT) and size of the imageability effect (mean RT low imageability minus mean RT for high imageability). Each point in the scatterplot represents a single individual. There is a significant correlation between the two measures (see table 8). In other words, the slower one is overall at performing synonym judgements the larger the

difference between your speed of response to high and low imageability stimuli. This relationship is important as individuals with aphasia are often slower to respond on such tasks than unimpaired controls (although this is not always the case). This slowing can be caused by a number of factors including the effects of age, brain damage, depression. Whatever the reason, clearly interpreting what is 'normal' needs to take this factor into account, using the scatterplot as a guide. Table 8 shows that there is a significant relationship between effect size and overall response speed for several other of the tasks (Visual Lexical Decision: frequency and imageability effects; Homophone Judgements: regular vs nonwords; exception vs nonwords; Scatterplots shown in Appendix). However, it is not the case that effect size correlated with overall speed of response for all tasks (Homophone judgements: regular vs exception; Rhyme judgements: rhyme and visual similarity effects).

Comparison of an individual to a (small) group of controls.

Thus far the message seems somewhat negative – control performance is variable and interpreting the performance of an individual person with aphasia is hence far from straightforward. However, there have been some statistical methods proposed that assist in this interpretation – providing us with estimates of an individual's 'abnormality' and confidence limits on these estimates. Crawford & Howell (1998) present a technique for comparing an individual's score to that of a small group of controls (modified independent samples *t*-test rather than z-score as is more usual for normative data from a large sample). This technique would be appropriate for the tasks reported here, to establish whether an individual showed speed or accuracy of performance that is significantly different from the control group. Crawford, Howell & Garthwaite (1998) extend the analysis to allow comparison of the difference between performance on two tasks (using modified paired samples *t*-test). For the tasks presented here this analysis would be appropriate for establishing whether the difference between speed (or accuracy) of two conditions was within the norm. For example, comparing whether the difference between high and low frequency stimuli on lexical decision, or between high and low imageability stimuli in synonym judgements. Crawford & Garthwaite (2002) extend these methods further and incorporate an estimation of the confidence limits of the results[3]. This allows an estimation not only of what proportion of the normal population would score lower (or respond slower) on a task, but also what the upper confidence limits

are on this estimation. These statistical tools help in the comparison of single cases to groups of control participants, although problems still remain by virtie of the variability in the normal population.

Summary

We have presented data from young Australian control participants performing four reading tasks from PALPA. The data from these young non-aphasic participants has confirmed that for some of these tasks:

- Ceiling effects in accuracy mask effects of psycholinguistic variables on normal performance that become apparent when speed of response is considered.
- The assumption of PALPA's creators that performance will be close to ceiling in accuracy is clearly erroneous for some of these tasks. Indeed, some participants perform close to chance on some conditions.
- Comparison of details of the pattern of performance for individual participants with that of a group of controls can be problematic given the variability within the controls. However, for at least some tasks there are reliable patterns of performance across individual controls.

These data will provide essential further information for clinicians and researchers alike when interpreting performance on these four PALPA subtests, and reinforce the importance of evaluating performance in terms of both speed and accuracy.

References

Best, W.M. (2000). Category-specific semantic disorders. . In W.Best, K.Bryan & J.Maxim (Eds.) *Semantic Processing in theory and practice.* London: Whurr.

Basso, A. (1996). PALPA: An appreciation and a few criticisms. Aphasiology, 10, 190-193.

Crawford, J. R., & Garthwaite, P.H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia, 40*, 1196-1208.

Crawford, J. R., & Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and observed scores. *Journal of Clinical and Experimental Neuropsychology*, *20*, 755-762.

Crawford, J. R., Howell, D.C., & Garthwaite, P.H. (1998). Payne and Jones revisited: Estimating the abnormality of test score differences using a modified paired samples t-test. *Journal of Clinical and Experimental Neuropsychology*, 20, 898-905.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. Behavior Research Methods, Instruments & Computers, 35, 116-124.

Kay, J., Lesser, R., Coltheart, M. (1992). PALPA: Psycholinguistic Assessments of Language Processing in Aphasia. Lawrence Erlbaum Associates, Hove, UK.

Kay, J., Lesser, R., Coltheart, M. (1996). PALPA: The proof of the pudding is in the eating. Aphasiology, 10, 202-215.

Marshall, J. (1996). The PALPA: A commentary and consideration of the clinical implications. Aphasiology, 10, 197-202.

Nickels, L.A. & Cole-Virtue, J.C. (2004). Effects of imageability on lexical decision latency. Manuscript in preparation.

Wertz, R.T. (1996). The PALPA's proof is in the predicting. Aphasiology, 10, 180-190.

Appendix A: Comparisons of Rhyme Judgement subtests using t-tests.

By subject

Reaction Time

	no rhyme-nonvissim		rhyme-vissim		rhyme-nonvissim	
	t	**p**	**t**	**p**	**t**	**p**
no rhyme-vissim	2.957	0.009	-6.314	<.001	-5.82	<.001
no rhyme-nonvissim			-6.439	<.001	-3.702	0.002
rhyme-vissim					-2.85	0.012

Error

	no rhyme-nonvissim		rhyme-vissim		rhyme-nonvissim	
	t	**p**	**t**	**p**	**t**	**p**
no rhyme-vissim	3.128	0.006	-4.585	<.001	-3.226	0.005
no rhyme-nonvissim			-4.443	<.001	-1.484	0.157
rhyme-vissim					-2.46	0.026

By item

Reaction Time

	no rhyme-nonvissim		rhyme-vissim		rhyme-nonvissim	
	t	**p**	**t**	**p**	**t**	**p**
no rhyme-vissim	1.914	0.066	6.353	<.001	5.484	<.001
no rhyme-nonvissim			4.301	<.001	3.125	0.004
rhyme-vissim					-1.585	0.124

Error

	t	**p**	**t**	**p**	**t**	**p**
no rhyme-vissim	1.889	0.69	4.743	<.001	3.276	0.004
no rhyme-nonvissim			2.719	0.015	1.153	0.259
rhyme-vissim					-2.606	0.015

Appendices B-J

These appendices can be downloaded from

http://www.maccs.mq.edu.au/~lyndsey/papers/N&C-V_2004_Appendices.xls

Appendix B: Visual Lexical Decision Item Accuracy Data for word stimuli.

Appendix C: Visual Lexical Decision Accuracy Data for nonword stimuli.

Appendix D: Individual Participant Analyses for Visual Lexical Decision.

Appendix E: Synonym Judgements Item Accuracy Data.

Appendix F: Individual Participant Analyses for Synonym Judgements.

Appendix G: Homophone Judgements Item Accuracy Data.

Appendix H : Individual Participant Analyses for Homophone Judgements.

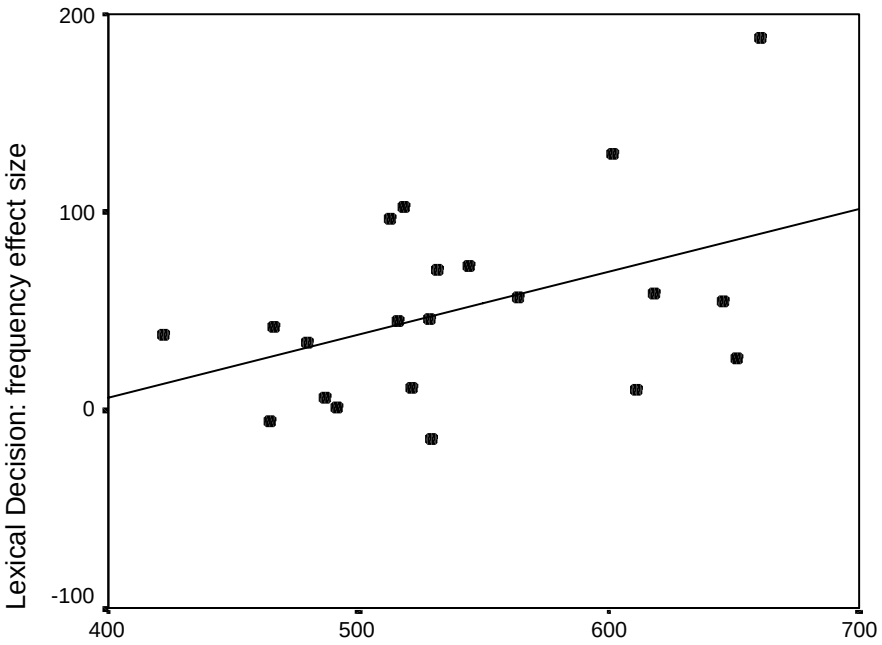
Appendix I: Rhyme Judgements Item Accuracy Data.

Appendix J: Individual Participant Analyses for Rhyme Judgements.

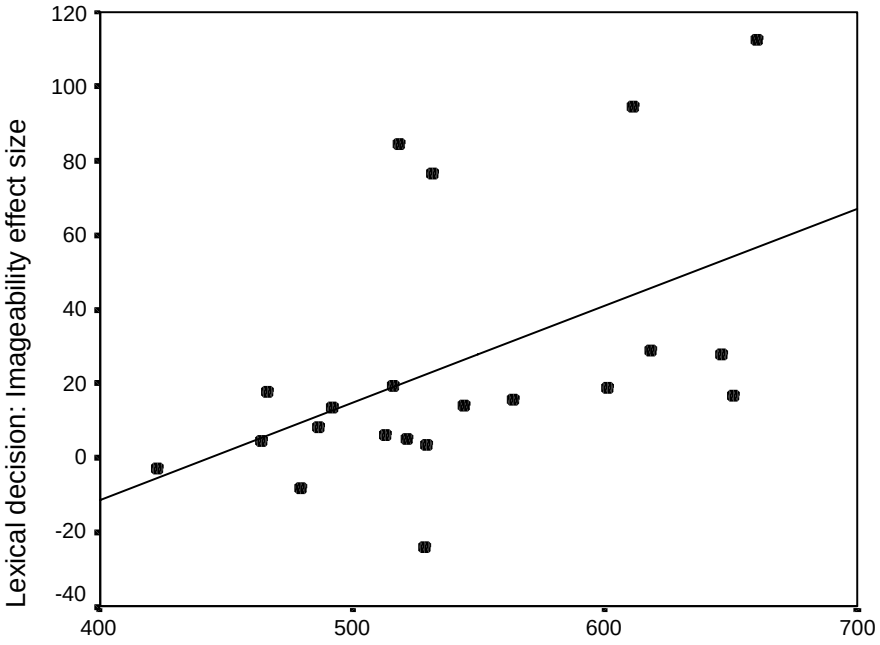
Appendix K

Scatterplots of the relationship between effect size and mean RT

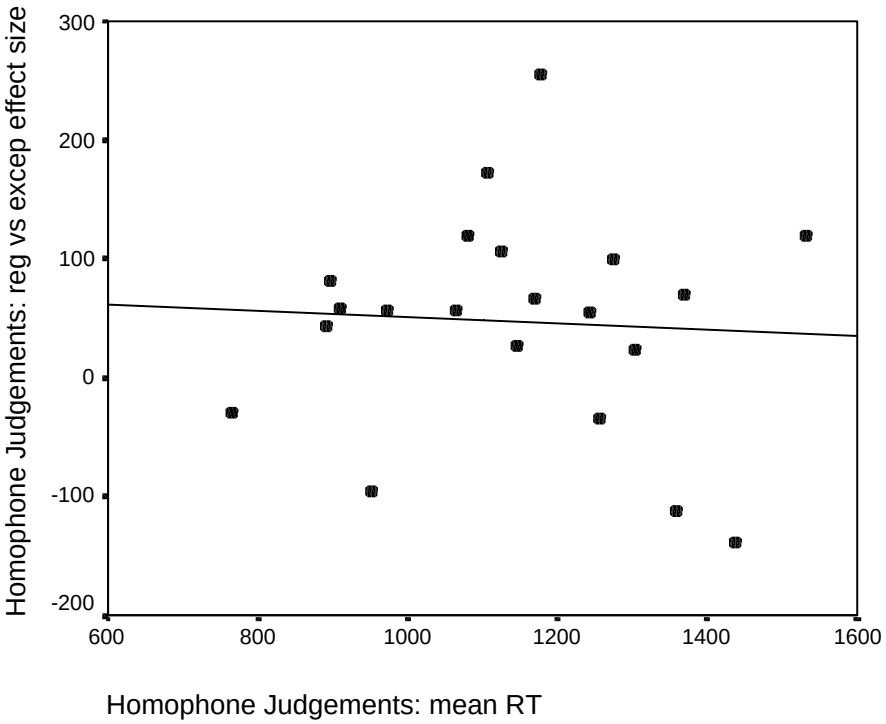
i) Lexical Decision & frequency



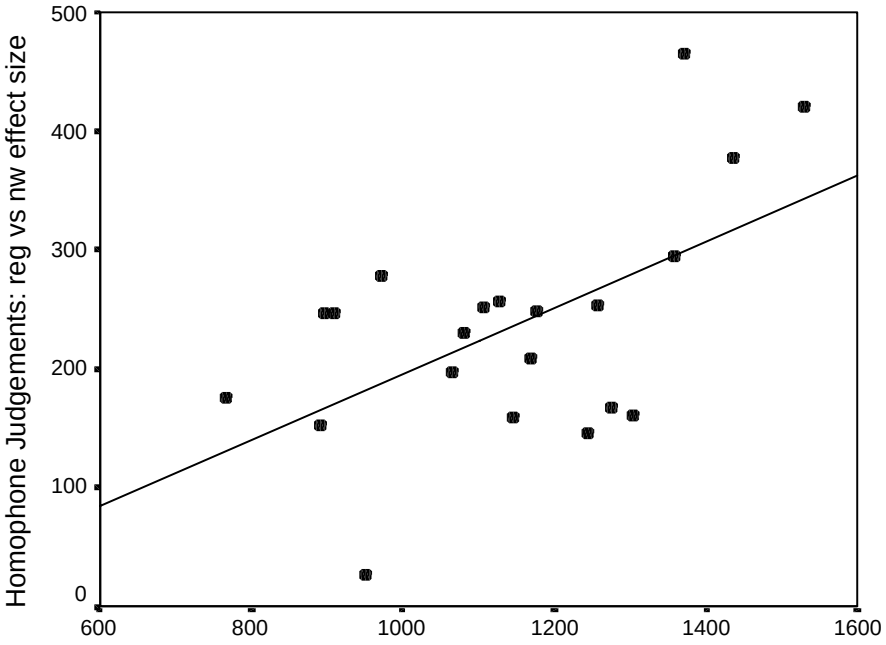
ii) Lexical Decision & Imageability



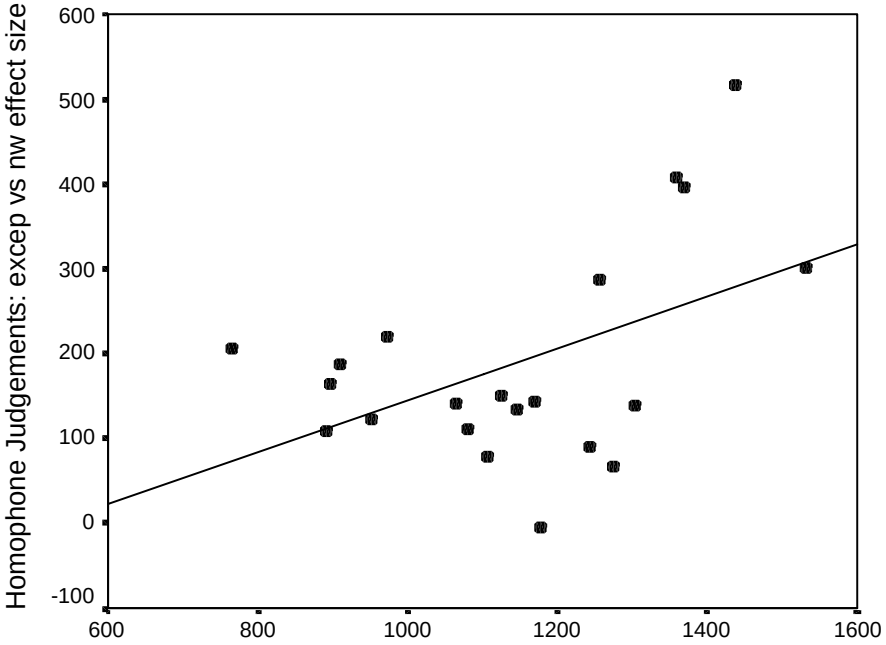
iii) Homophone Judgements: regular and exception words



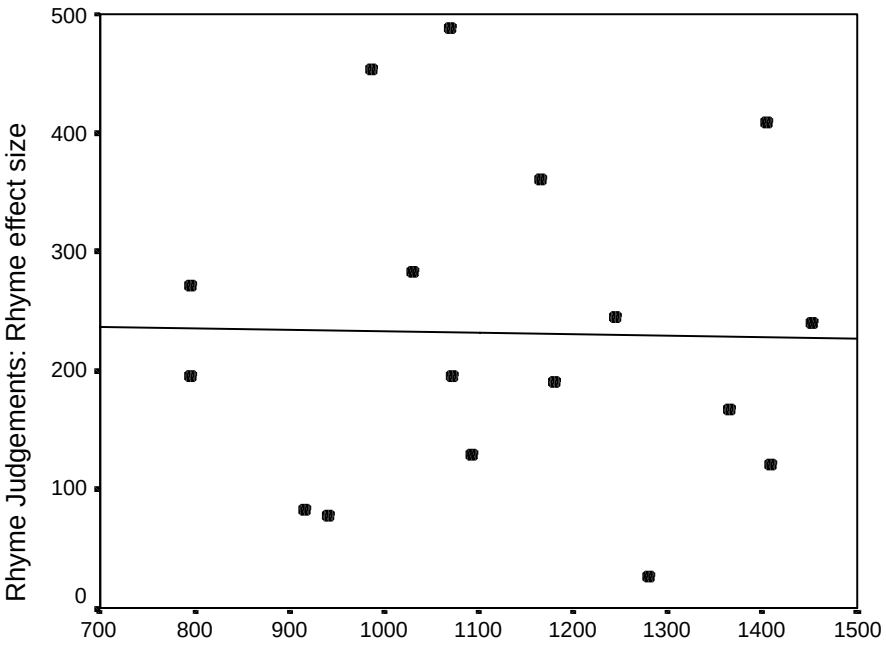
iv) Homophone Judgements: regular vs nonwords



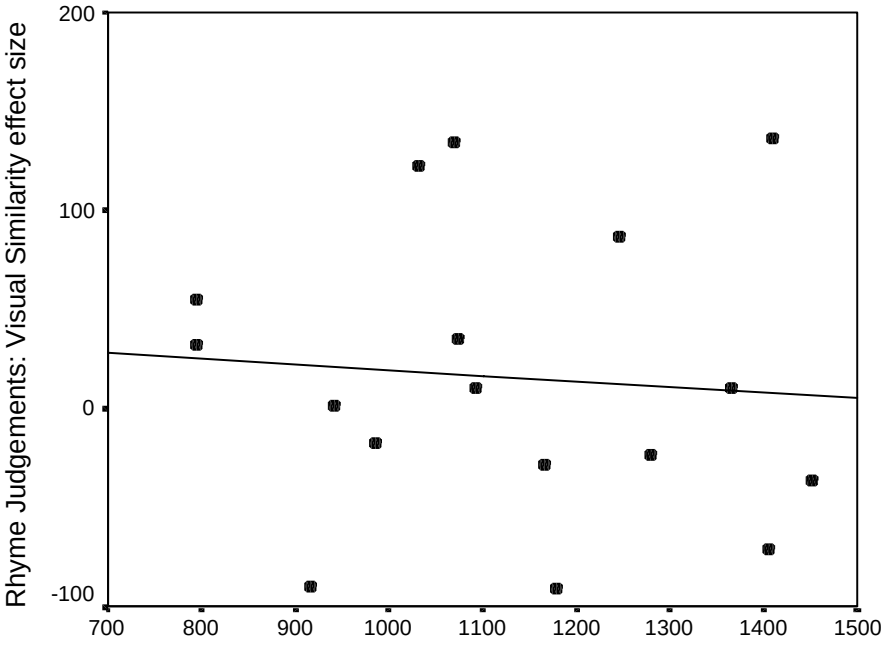
v) Homophone Judgements: exception vs nonwords



vi) Rhyme Judgements: Effect of rhyme



vii) Rhyme Judgements: Effects of visual similarity



Authors' Note

Lyndsey Nickels was supported by an Australian Research Council QEII fellowship during preparation of this paper. Thanks to Anna Woollams for programming the DMDX software, to Carl Windhorst for help running the experiments, and to Britta Biedermann for some of the analysis. Two reviewers provided helpful comments on an earlier draft and David Howard suggested the inclusion of effect sizes and their relationship with mean RT.

Footnotes

[1] The 120 stimuli from subtest 25 were presented together with the additional 40 stimuli that occur in Subtest 5 (auditory lexical decision); we only report the data from the subset of 120 items that is consistent with the items in subtest 25 (visual lexical decision).

[2] For performance to be significantly better than chance a score of 20 or more correct out of 30 is required, or 12 out of 15.

[3] Computer programmes for performing these calculations can be obtained are made available by John Crawford and are downloadable from http://www.abdn.ac.uk/~psy086/dept/abnolims.htm

Table 1: Latency and Accuracy Data for Visual Lexical Decision, by-subjects.
Note that the same 60 stimuli comprise the High and Low Imageability sets as the High and Low frequency sets (ie there are four subsets High Imageability, High frequency; High Imageability Low frequency; Low Imageability, High frequency; Low Imageability Low frequency)

	Imageability		Frequency		
Mean Reaction Time	High Imageability	Low Imageability	High Frequency	Low Frequency	Mean
Words (Yes Response)	528.34 (60.31)	553.74 (77.26)	516.33 (60.49)	567.40 (60.49)	541.06 (67.07)
Non-Words (No Response)	N/A	N/A	N/A	N/A	638.24 (94.33)
Mean					589.13 (76.60)

	Imageability		Frequency		
Mean Accuracy (SD)	High Imageability (n=30)	Low Imageability (n=30)	High Frequency (n=30)	Low Frequency (n=30)	Total (n=60)
Words: Yes Response	29.14 (0.91)	28.71 (1.52)	29.67 (0.58)	28.19 (1.69)	57.86 (1.74)
Non-Words: No Response	N/A	N/A	N/A	N/A	57.38 (2.75)
Total (n=120)					115.24 (4.21)

Table 2: Latency and Accuracy Data Synonym Judgements, by-subjects.

Mean Reaction Time (SD)	High Imageability Items	Low Imageability Items	Mean
Synonymous pairs (Yes Responses)	882.71 (172.28)	1141.82 (310.90)	1008.28 (231.54)
Non-synonymous pairs (No Responses)	980.02 (231.77)	1129.06 (260.29)	1053.22 (238.84)
Mean	931.76 (195.80)	1132.47 (267.97)	1030.17 (226.45)

Number correct (SD)	High Imageability (n=15)	Low Imageability (n=15)	Total Correct (n=30)
Synonymous pairs (Yes Responses)	14.15 (0.75)	13.40 (1.19)	27.55 (1.43)
Non-synonymous pairs (No Responses)	14.75 (0.55)	14.45 (1.15)	29.20 (1.58)
Total (n=30)	28.90 (0.85)	27.85 (1.69)	56.75 (2.15)

Table 3: Latency and Accuracy Data for Homophone Judgement, by-subjects.

Mean Reaction Time (SD)	Regular Words	Irregular Words	Nonwords	Mean
Homophonic pairs (Yes Responses)	955.03[1] (205.60)	1067.73[3] (213.32)	1265.34[5] (280.96)	1096.03 (213.85)
Non-homophonic pairs (No Responses)	1157.40[2] (214.31)	1131.90[4] (203.71)	1315.21[6] (237.71)	1201.50 (197.48)
Mean	1051.47 (189.28)	1098.66 (189.04)	1287.52 (247.10)	1144.26 (312.09)

Mean No. correct (SD)	Regular Words (n=10)	Irregular Words (n=10)	Nonwords (n=10)	Mean Total Correct (n=30)
Homophonic pairs (Yes Responses)	9.33[1] (0.48)	9.24[3] (0.70)	9.19[5] (1.03)	27.76 (1.04)
Non-homophonic pairs (No Responses)	9.14[2] (1.20)	9.24[4] (0.89)	8.67[6] (1.20)	27.05 (2.67)
All Items (n=20)	18.48 (0.93)	18.48 (1.12)	17.86 (1.68)	54.81 (2.77)

1 R: Regular;
2 RC: Regular Control;
3 E: Exception;
4 EC: Exception Control;
5 NW: Nonword;
6 NWC: Non-word Control

Table 4: T-tests of Latency and accuracy Data for Word Types in Homophone Judgement, by-subjects and by-items.

	By-Subjects				By-Items			
Word Type	Mean RT		Accuracy		Mean RT		Accuracy	
	t	p	t	p	t	p	t	p
Regular vs. irregular	2.349	0.029	1.520	0.144	1.108	0.275	.000	1.000
Regular vs. nonwords	10.986	0.000	-.780	0.444	5.348	0.000	.882	0.384
Irregular vs. Nonwords	6.776	0.000	-.322	0.751	5.135	0.000	1.040	0.305

Table 5: Latency and Accuracy Data for Rhyme Judgement, by-subjects.

Mean Reaction Time (SD)	Rhyme	Non-Rhyme	Mean
Visually Similar	984.15[1] (217.80)	1321.85[3] (263.94)	1121.24 (215.66)
Non-Visually Similar	1069.87[2] (229.81)	1206.45[4] (210.69)	1136.84 (209.53)
Mean	1025.53 (214.27)	1257.11 (219.10)	1129.47 (208.71)

Mean Accuracy (SD)	Rhyme (n=15)	Non-Rhyme (n=15)	Total (n=30)
Visually Similar	0.353[1] (0.493)	3.47[3] (2.98)	26.18 (3.23)
Non-Visually Similar	1.18[2] (1.24)	1.88[4] (1.78)	26.94 (2.28)
Total (n=30)	28.47 (1.28)	24.65 (4.40)	53.12 (5.10)

1 SPR: Spelling Pattern Rhyme;
2 PR: Phonological Rhyme;
3 SPC: Spelling Pattern Control;
4 PC: Phonological Control

Table 6: Summary of overall mean reaction time, and accuracy for four PALPA tasks, with values for a cut off of two standard deviations below the mean for each measure.

			Reaction Time			Number correct			Number of errors			Number of control participants
Subtest No.		n	mean	SD	2SD below	Mean	SD	2 SD below	Mean	SD	2 SD below	
25	Visual Lexical decision	120	589.13	76.60	742.43	115.24	4.21	106.83	4.76	4.21	13.17	21
50	Synonym Judgements	60	1030.17	226.45	1483.08	56.75	2.15	52.45	3.25	2.15	7.55	20
28	Homophone Judgements	60	1144.26	312.09	1768.45	54.81	2.77	**49. 27 44.43	5.19	2.77	10.73	21
15	Rhyme Judgements	60	1129.47	208.71	1546.9	53.12	5.1	42.92	6.88	5.1	17.08	17

Table 7: Summary of those factors that significantly affected young control participant performance on four PALPA reading tasks.

Legend:
* By subjects only
Note as accuracy is often at ceiling, examination of effects on individuals was often not appropriate, see text for further discussion.
1: All effects significant in the same direction as the group results

		Group effects significant for..		**% of Individuals that show significant effects[1]**		**% of Individuals that show effects in the same direction as the group**		**Effects reliable for all individual subjects**	
Task	**Variable**	**Reaction Time**	**Accuracy**	**RT**	**Accuracy**	**RT**	**Accuracy**	**RT**	**Accuracy**
Visual Lexical Decision	Frequency	✓	✓	38%	5%	90%	95%		
	Imageability	✓	×	14%	5%	86%	76%		
	Lexicality	✓	×	-	-	-	-		
Synonym Judgements	Imageability	✓	×	70%	-	100%	80%	✓	
	Synonymy	×	✓	-	-	-	-		
Homophone Judgements	Word Type	✓	×		-	71%	-		
	Reg vs. Exception	✓*	×	10%	-	76%	-		
	Reg vs. Nonwords	✓	×	71%	-	100%	-	✓	
	Exception vs. Nonwords	✓	×	43%	-	95%	-		
	Homophony	✓	×	-	-	-	-		
Rhyme Judgements	Rhyme	✓	✓	53%	29%	100%	94%		
	Visual Similarity	×	×	0%	0%	-	-		
	Rhyme*Visual Similarity	✓	✓	24%	-	65%			

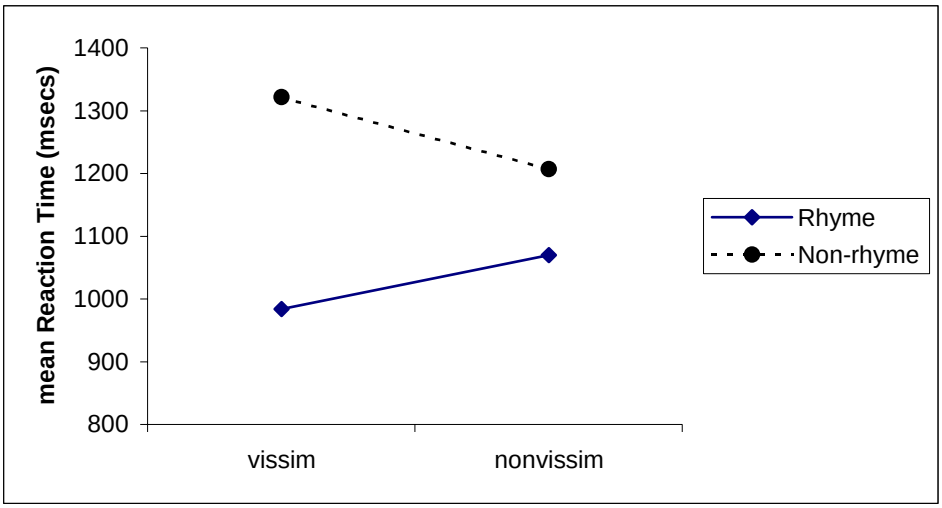
Table 8: Mean effect sizes (RT), 95% Confidence Intervals and correlation of effect size and Mean RT.

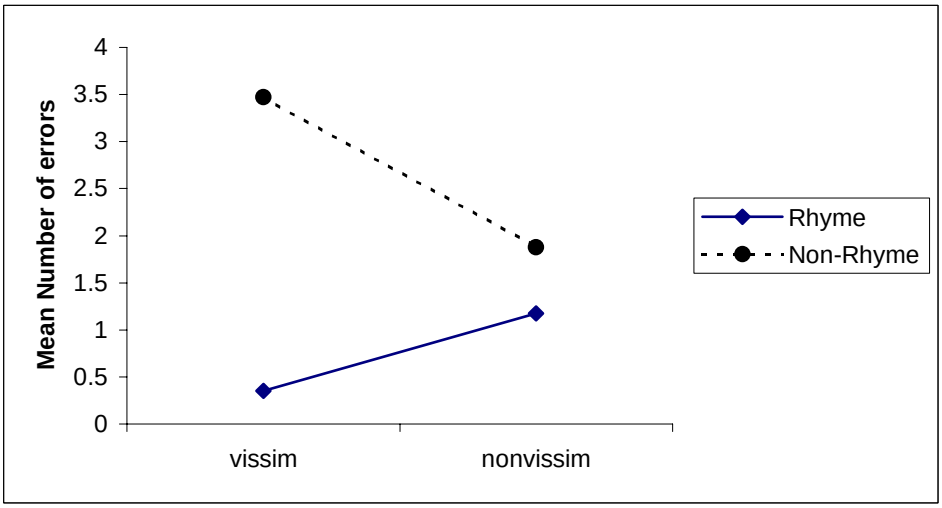
				95% confidence intervals		Correlation with mean RT	
Task	Variable	Direction of Calculation	Mean effect size (msecs)	Upper	Lower	r	p=
Visual Lexical Decision	Frequency	Low Freq - High freq	51.1	71.9	30.3	0.436	0.048
	Imageability	Low Image- High Image	25.4	40.7	10.1	0.489	0.024
Synonym Judgements	Imageability	Low Image - High Image	200.7	248.7	152.7	0.668	0.001
Homophone Judgements	Reg vs Exception	Exception - regular	47.5	87.1	7.9	-0.059	ns
	Reg vs Nonwords	Nonwords - regular	236.4	278.5	194.2	0.566	0.008
	Exception vs Nonwords	Nonwords- exception	188.9	243.5	134.2	0.480	0.028
Rhyme Judgements	Rhyme	Nonrhyme-rhyme	231.6	295.2	168.0	-0.020	ns
	Visual Similarity	Nonvissim-vissim	15.6	50.3	-19.1	-0.082	ns

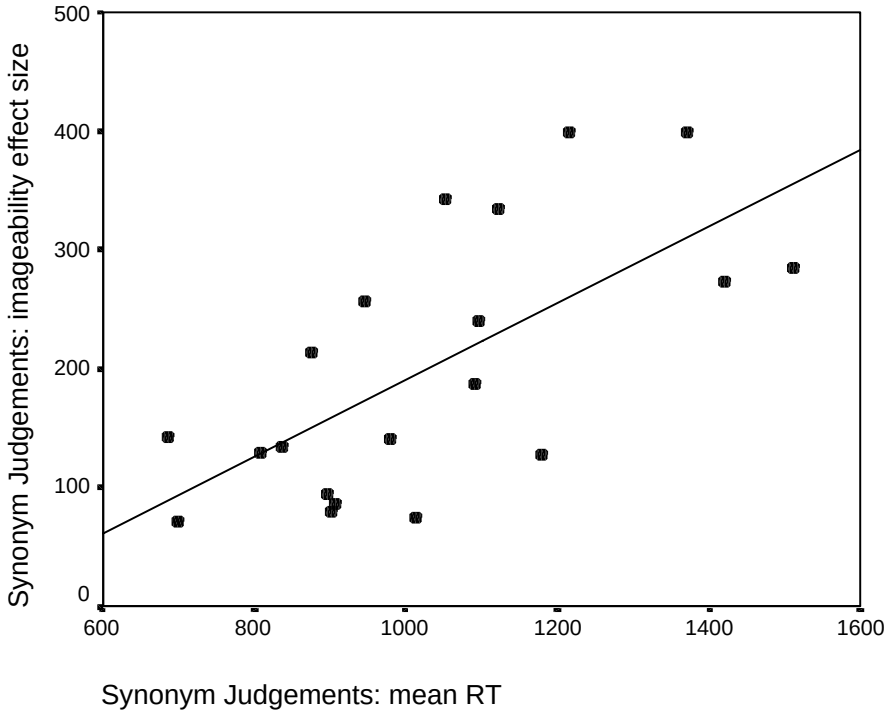
Figure Legend

Figure 1: Interaction between effects of rhyme and visual similarity on reaction time and error rate in the rhyme judgements task.

Figure 2: Scatterplot of the relationship between size of the imageability effect (reaction time difference) and overall mean reaction time for Synonym Judgements.


mean Reaction Time (msecs)
1400
1300
1200
1100
1000
900
800
vissim
nonvissim
Rhyme
Non-rhyme


Mean Number of errors
4
3.5
3
2.5
2
1.5
1
0.5
0
vissim
nonvissim
Rhyme
Non-Rhyme


Synonym Judgements: imageability effect size
500
400
300
200
100
0
600
800
1000
1200
1400
1600
Synonym Judgements: mean RT