Running head: AUDITORY LEXICAL DECISIONS IN DLD

Auditory lexical decisions in developmental language disorder: A meta-analysis of behavioural studies

Samuel David Jones and Silke Brandt

Lancaster University

Correspondence concerning this article should be addressed to Sam Jones, Department of Linguistics and English Language, County South, Lancaster University, Lancaster, UK, LA1 4YL. Email: sam.jones@lancs.ac.uk

Abstract

Purpose: Despite the apparent primacy of syntactic deficits, children with developmental language disorder (DLD) often also evidence lexical impairments. In particular, it has been argued that this population have difficulty forming lexical representations that are detailed enough to support effective spoken word processing. In order to better understand this deficit, a meta-analysis of studies testing children with DLD in the auditory lexical decision task was conducted. The objective was to provide summary effect size estimates for accuracy and response time measures, for comparisons to age- and language-matched control groups.

Method: Two thousand three hundred and seventy-two (2372) records were initially identified through electronic searches and expert consultation, with this cohort reduced to nine through duplicate removal and the application of eligibility and quality criteria. The final study cohort included 499 children aged 3;8-11;4.

Results: Multivariate analysis suggests that children with DLD were significantly less accurate in the auditory lexical decision task than age-matched controls. For the response time estimate, however, confidence intervals for the same group comparison crossed zero, suggesting no reliable difference between groups. Confidence intervals also crossed zero for language-matched control estimates for both accuracy and response time, suggesting no reliable difference between groups on either measure.

Conclusion: Results broadly support the hypothesis that children with DLD have difficulty forming detailed lexical representations relative to age- though not language-matched peers. However, further work is required to determine the performance profiles of potential subgroups and the impact of manipulating different lexical characteristics, such as the position and degree of non-word error, phonotactic probability, and semantic network size.

Keywords: Developmental language disorder (DLD); specific language impairment (SLI); auditory lexical decision task; meta-analysis.

Introduction

Children with developmental language disorder (DLD; also specific language impairment, or SLI), show severe language deficits in the absence of frank neurological damage, acquired epileptic aphasia, autism-like behavior, sensory-neural hearing loss, or genetic conditions such as Down syndrome or cerebral palsy (Bishop, Snowling, Thompson, & Greenhalgh, 2016). While morpho-syntactic deficits are the hallmark of DLD, spoken word processing is also commonly impaired (see Kan & Windsor, 2010, for review). Affected children may, for instance, have difficulty repeating non-words accurately (Graf Estes, Evans, & Else-Quest, 2007), or may require longer auditory strings than age-matched controls in order to recognise a word in the gating paradigm (e.g. Dollaghan, 1998; Montgomery, 1999).

The current meta-analysis looks at the auditory lexical decision task, in which participants are required to provide a 'yes'/'no' or non-linguistic (i.e. button press) judgement response to auditory word and non-word stimuli. For instance, in response to the word *dinosaur* (/daɪnəsɔːr/) the participant is required to make an affirmative response, while in response to the non-word *dinokor* (/daɪnəkɔːr/) the participant is required to reject the stimulus. Accuracy and response time may be recorded, and word and non-word stimuli are normally manipulated in line with primary research aims. This may include, for instance, controlling target word frequency, phonotactic probability, the number of semantically associated words (i.e. semantic network size), the position of non-word error, e.g. *dinokor* (/daɪnəkɔːr/) versus *kinosaur* (/kaɪnəsɔːr/), and the degree of non-word divergence, e.g. *dinokor* (/daɪnəkɔːr/) versus *kinokor* (/kaɪnəkɔːr/).

In its conventional form, the auditory lexical decision task is argued to measure 'the quality or precision of stored phonological representations at the whole-word level' (Claessen & Leitão, 2012, p. 215), with accurate rejection of a non-word taken as evidence that the

corresponding word-level, phonological representation is appropriately detailed. As such, the lexical decision paradigm constitutes a useful tool to examine the hypothesis that children with DLD have difficulty forming detailed lexical representations in long-term memory, potentially as a result of underlying auditory processing or short-term memory deficits (Bishop, 1997). This pattern of development constitutes a delay rather than deviance, with young, typically developing children also apparently forming relatively holistic lexical representations prior to the emergence of a system of phonemic representation that supports the retention, and accurate and rapid processing of minimally different words (e.g. /kæt/ and /kæʃ/); a transition interacting closely with growth of the lexicon (Walley, 1993; see, however, Ainsworth, Welbourne, & Hesketh, 2016, for an interpretation of early underspecification-like performance in terms of the complexity of task demands).

In this context, the auditory lexical decision task has a number of advantages over other paradigms. First, the task arguably resembles natural spoken word recognition more closely than alternatives such as gating or non-word repetition, and so results may be more generalisable. Second, in requiring only a button touch or minimal verbal response, the task minimises the possibility that performance deficits stem from the motor output level rather than underspecification of the lexicon; an interpretation not ruled out by paradigms requiring more complex verbal responses, for instance naming and non-word repetition.

Superficially, there may be little question that children with DLD perform worse than age-matched controls on the auditory lexical decision task. However, previous meta-analyses of associated paradigms (e.g. non-word repetition; Graf Estes et al., 2007) indicate that there may exist heterogeneity in effect sizes that is masked by a general emphasis on statistical significance. The meta-analytic approach facilitates the fine-grained assessment of such heterogeneity, enabling researchers to examine which particular clinical profiles or task design features are associated with smaller or larger performance discrepancies. In doing so,

results may improve our understanding of factors inhibiting spoken word processing in this population, and provide a platform for the development of evidence-based practice. Better understanding of this deficit is important because protracted lexical underspecification may have a detrimental impact on various areas of linguistic development and behaviour, including not only vocabulary learning and spoken word recognition and production, but also grammatical development and literacy (Claessen & Leitão, 2012; Goodman & Bates, 1997).

Given the extensive use of the lexical decision task in clinical and non-clinical contexts, this report may be of interest to both researchers and practitioners. The population effect size estimates may provide a useful benchmark for future research, for instance when conducting prospective power analyses or for researchers adopting a Bayesian analytical framework in which parameters must be specified apriori. Data aggregation is particularly valuable in the field of DLD given the prevalence of studies with low sample sizes, often entailing low statistical power and a high false positive rate, i.e. small samples are more likely to produce extreme values (Robey & Dalebout, 1998). The question examined is:

> *What are the estimated population effect sizes of the discrepancies in performance (response accuracy and latency) between children with DLD and age- and language-matched controls on the auditory lexical decision task?*

A substantial literature documenting lexical processing deficits across a range of paradigms (see Kan & Windsor, 2010) suggests population estimates will indicate age-matched controls regularly outperform children with DLD, with higher accuracy rates and lower response times. However, given that evidence of lexical underspecification is held to reflect a developmental delay rather than deviance (Bishop, 1997), it may be reasonable to expect little difference in estimates between children with DLD and language-matched controls.

Method

This study was pre-registered with the Open Science Framework on June 9th, 2017, with a protocol available from the associated project page (see https://osf.io/2cvnm/). The study fulfils Preferred Reporting Items for Systematic reviews and Meta-Analysis guidelines (PRISMA, see http://prisma-statement.org), with a completed checklist also available from the Open Science Framework project page.

Eligibility criteria

Participants. The population of primary interest was atypically developing children and adolescents, defined as those prior to or in full-time education, with age- and language-matched control groups included on the basis of provision in primary studies. Atypically developing was defined as children with DLD, as described by Bishop et al. (2016) and repeated in the introduction to the current study. A summary of the CATALISE statement on diagnostic terminology can be found at: https://naplic.org.uk/sites/default/files/Summary%20of%20CATALISE%20%28v3%29.pdf. Participants were not distinguished on the basis of age, gender, socio-economic status, ethnicity, language, or geographical location.

Experimental design. Studies of interest were those using the auditory lexical decision task to test children with DLD. Studies were required to use experimental and control groups. No single-subject case studies were included, though there was no lower boundary on cohort size.

Outcome measures. The values of interest were the means and associated standard deviations of typical and atypical group performance on the auditory lexical decision task. This could be an accuracy rate (percentage or raw score) and/or a response time (in milliseconds; note that response times are typically only included for accurate responses in the primary literature). Standardised mean differences and variances were calculated from these primary statistics, in addition to group sizes. Throughout this study, negative effect

sizes for accuracy outcomes indicate that children with DLD were less accurate than controls, while positive effect sizes for response time indicate that children with DLD were slower to respond.

Types of study. Journal articles, research reports, book chapters, and grey literature, including conference abstracts and unpublished theses and datasets were considered for inclusion. Accommodating grey literature is crucial to mitigating the impact of publication bias, whereby significant results are more likely to be published than non-significant results. Newspapers, magazine articles, and blogs were excluded. There was no restriction on the date of publication.

Defining and piloting search terms

Initial scope searches using the free text strings *specific language impairment, developmental language disorder,* and *lexical decision* were conducted on June 1st, 2017, using the databases PubMed, PsychINFO, Web of Knowledge, and Linguistics and Language Behavior Abstracts. These searches returned eleven studies testing clinical populations using the auditory lexical decision task, from which specific search terms were extracted from keyword lists (see Table 1).

[Table 1]

These initial scope searches revealed that the paradigm was referred to variously as the auditory lexical decision task and the auditory lexical judgement task. In addition, there were anticipated differences between diagnostic labels, prominently: SLI, DLD, and language impairment. Main search terms were defined to accommodate this diversity. In particular, a strategy was developed using Boolean operators to link variations in diagnostic terminology to variations in paradigm terminology. An example search strategy in simplified (i.e. no field specification or MeSH terms) PubMed format is:

```
(specific language impairment OR developmental language disorder
OR language impairment) AND (auditory lexical decision OR
auditory lexical judgement)
```

Piloting this strategy on PubMed on June 5th, 2017 returned 67 results. The number of records retrieved did not increase with the inclusion of alternative diagnostic labels including primary language impairment, developmental dysphasia, or language disability. Note that none of the finalised search terms listed above differ in British and American English spelling.

Main search strategy

Four approaches were used in evidence gathering: Electronic database searches, journal searches, bibliographic searches, and expert consultation. First, the following seven electronic databases were searched using the strategy specified above: Scopus, PubMed, Web of Science, LLBA, JSTOR, OVID, and ERIC. Second, forty-six journals in child language, psycholinguistics, speech-language therapy, and developmental psychology were hand searched using the aforementioned search terms and associated free text strings. The journals examined were identified during prior electronic database searches, and are listed in full on the Open Science Framework page associated with this project (see https://osf.io/2cvnm/). Third, the literature reviews and reference sections of retrieved papers were hand searched for further relevant papers. Fourth, 55 researchers were contacted regarding overlooked studies and the availability of unpublished datasets. The email sent included a link to the pre-registration protocol and a spreadsheet of studies retrieved prior to consultation, specifying author, year, title, and DOI with hyperlinks to the primary sources. The pre-registered stop search date for all data gathering was August 29th, 2017.

Quality, strength of evidence, and bias risk assessment

The strength of the body of evidence collected using these four search strategies was assessed according to the following criteria. Only papers that met these criteria were included; there was no ranked quality index.

1. Studies lacking an appropriate control group or data required to compute standardised mean differences and sampling variances (e.g. *M*, *SD, N/n*) after author consultation were excluded.
2. Studies lacking primary diagnostic or linguistic data (e.g. age, non-verbal IQ, standardised language test scores) for experimental or control groups were excluded.
3. Studies in which authors declared conflict of interest were excluded.

Statistics from studies meeting the above quality criteria were extracted for inclusion in the meta-analysis.

Meta-analysis

Data extraction. Studies were attributed numeric IDs and coded by: (a) author(s); (b) year of publication; (c) DLD group mean chronological age; (d) DLD group mean language age; (e) control type (i.e. age- or language-matched); (f) outcome measure (i.e. accuracy or response time); (g) stimulus type (i.e. words, non-words); (h) stimuli sub-classification, commonly unique to the aims of original study (e.g. word initial or final manipulation in non-word formation); (i) mean scores (typically a percentage for accuracy outcomes, with response times specified in milliseconds), standard deviations, and sample sizes of DLD groups; and (j) mean scores, standard deviations, and sample sizes of control groups. Coding was conducted by the first author, with a random sample of five studies then repeated by a trained coder. Disagreements were resolved through re-examination until agreement was 100%. The complete dataset can be built using the R code available from the Open Science Framework page associated with this project (see https://osf.io/2cvnm/).

Package and model selection. The meta-analysis was conducted using the Metafor package in R (Viechtbauer, 2010). This package was chosen because it is freely available and the associated code can be easily disseminated in the interests of quality assessment and replication. Metafor is also able to manage complex datasets like that analysed in the current study, with multiple control groups and dependent measurements. Given that a number of studies include both accuracy and latency outcomes (i.e. multiple-endpoints; Gleser & Olkin, 2009), as well as two types of control group (age- and language-matched), the decision was taken to fit a multivariate, random-effects model, which would accommodate stochastically dependent effect sizes while providing an overall estimate for each control group and outcome pairing.

Procedure. With the data frame in R, standardised mean differences (Hedges' *g*: Hedges, 1981) and sampling variances were computed using `metafor::esclac()`. In the current study, there are four comparisons of interest (see Table 2).

[Table 2]

In order to retrieve estimates for each of these combinations (i.e. groups (age-matched, language-matched) with outcomes (accuracy, response time)), dummy variables were created and plugged into the linear model specified within the `rma.mv()` function as moderators. The model was then passed to the `robust.rma.mv()` function, which provides a robust estimate of the variance-covariance matrix of model estimates and computes tests and confidence intervals of coefficients using a small-sample adjustment. Adopting the same procedure, two additional models were fitted which specified identical moderators plus random effects at (a) study level (denoted 'author'; see model 2), and (b) both study and outcome levels (i.e. accuracy and response time; see model 3). This reflects the assumption that the underlying true effects within these levels will be more similar than the underlying true effects from different levels (see http://www.metafor-project.org/doku.php/analyses:

konstantopoulos2011). Model fit was then compared using `fitstats()` to retrieve Akaike information criterion values, before identifying potential outliers calculating standardised residuals; `rstandard()`. Publication bias risk was assessed using fail-safe N, which provides an estimate of the number of additional studies reporting negligible effects required in order to nullify a summary effect (Rosenthal, 1979; Orwin, 1983; Rosenberg, 2005). If this number is relatively large, it may be inferred that the estimate is unlikely to be compromised by publication bias.

Search results and study selection

Figure 1 shows the number of studies retrieved through searches and expert consultation, and the number excluded during preliminary screening and quality and eligibility assessment. A total of 2340 records were retrieved through electronic database searches. A full record of our electronic database searches is available from the Open Science Framework page associated with this project (see https://osf.io/2cvnm/). Twelve unique records were then retrieved through bibliographic and hand searches using the aforementioned search terms and associated free text strings. The response rate to expert consultation emails was 20%, with eleven contributing author comments received and twenty studies not previously identified recommended for inclusion. In all, 2372 records were retrieved, with 2335 then excluded through duplicate removal and the screening of abstracts in line with the aforementioned criteria. This brought the number of studies sent to full-text quality and eligibility assessment to thirty-seven. At this stage the cohort included four articles considered grey literature: One pre-print, one poster, one doctoral thesis, and one research report. The bottom right panel of Figure 1 lists the rationales for excluding 28 studies during full-text appraisal and quality assessment. Nine studies were ultimately included in the meta-analysis, all of which were published in peer-reviewed journals between 1994 and 2016. No contributing authors declared potential conflict of interest.

[Figure 1]

Figure 1. Study search and selection flow diagram.

Description of selected studies

An extensive summary of the nine studies included in the meta-analysis is presented in Appendix A, which details: (a) author, (b) year, (c) type(s) (i.e. age- or language-matched), ages, and sample sizes of experimental and control groups, (d) the standardised tests used to determine DLD, age-matched, and language-matched control groups[1], (e) stimulus type and number (including sub-classifications), (f) response type (verbal or non-verbal), and (g) outcome measure (accuracy or response time).

Participants. The nine studies involved a total of 499 participants: 191 with DLD (age range 3;8-11;4), 120 age-matched control participants (age range 7;3-11;4), and 188 language-matched control participants (age range 4;1-9;7). One study included only age-matched controls, while three studies included only language-matched controls. The remaining five studies included both age- and language-matched controls. Five studies included participants whose first language was English, while three tested French-speaking children, and one tested Brazilian-Portuguese speakers. Five studies specified that participants were monolingual, with monolingual or multilingual status unclear in the remaining studies.

The diagnostic criteria used in each study are specified in Appendix A. Diagnosis commonly worked on the basis of a verbal/non-verbal IQ discrepancy. In two studies (Edwards & Lahey, 1996; Windsor & Hwang, 1999) data from standardised diagnostic tests was use to identify subgroups, namely expressive-only (termed SLI-expressive) and

[1] Note that the standardised measures used to determine language-matched control groups differed between studies (see Appendix A).

expressive-receptive DLD (termed SLI-mix). In one study (Crosbie, Howard, & Dodd, 2004), subgroups not initially identified through standardised tests were defined post-hoc on the basis of auditory lexical decision task data. Befi-Lopes, Pereira, and Bento (2010) and Maillart, Schelstraete, and Hupet (2004) also include language-impaired, lexical-age subgroups based on receptive vocabulary test performance. The remaining studies did not differentiate subgroups. Note, relatedly, that the test group of James, Van Steenbrugge, and Chiveralls (1994) comprised language-disordered children with concurrent central auditory processing deficits.

Results

Meta-analysis

Three robust, multivariate, random effects models were fitted. Model one specified moderators only (i.e. dummy variables specifying comparison and outcome pairing; see Table 2); model two specified moderators and random effects at study level; and model three specified moderators and random effects at both study and outcome levels. These models were compared using Akaike information criterion (AIC); a parsimony-adjusted measure of relative model fit based on out-of-sample deviance (McElreath, 2016). The results of this process are summarised in Table 3.

[Table 3]

Decreases in AIC indicate that model fit is improved considerably by the specification of random effects at study level, and marginally by the additional specification of random effects at outcome level. Computing internally standardised residuals for model three suggested no significant outliers according to a +/-2 threshold (though a small number of cases approached this figure; see R code). Accordingly the estimates and confidence intervals reported here are derived from model three. Table 4 presents a full model summary.

[Table 4]

Four primary observations can be taken from the output shown in Table 4. First, children with DLD are substantially less accurate than age-matched controls in the lexical decision task, with an estimated Hedges' *g* of -0.88 (SE=0.19) and a confidence interval bound below zero (95%CI = -1.37 to -0.39), suggesting a robust population effect. Second, despite a moderate effect size, the confidence interval for the estimate reflecting response time discrepancies between children with DLD and age-matched controls marginally crosses zero (Hedges' *g* = 0.53; SE=0.21; 95%CI = -0.01 to 1.06), suggesting zero or values approaching zero are a reasonable possibility for the population effect. Third, the confidence interval for the moderate effect size reflecting accuracy discrepancies between children with DLD and language-matched controls crosses zero (Hedges' *g* = -0.46; SE=0.23; 95%CI = -1.06 to 0.13). Fourth, the confidence interval for the small effect size reflecting response time discrepancies between children with DLD and language-matched controls crosses zero (Hedges' *g* = 0.22; SE=0.13; 95%CI = -0.10 to 0.54). Forest plots visualising case and summary effect sizes and confidence intervals grouped by control type and outcome measure are presented in Appendix B. These plots indicate considerable variability between both studies and cases. For instance, Figure B3 shows differing estimates for cases 87 (Hedges' *g* = -0.26; 95%CI = -0.78 to 0.26) and 88 (Hedges' *g* = 0.15; 95%CI = -0.36 to 0.67), both of which record discrepancies in accuracy judgements to real words between children with DLD and language-matched controls in the Haebig, Kaushanskaya, and Ellis Weismer (2015) study. Two factors potentially contributing to between- and within-study variability, namely sample heterogeneity and differences in the manipulation of experimental stimuli, are considered at length in the discussion section, where we also justify our decision not to include stimuli sub-classifications or posited subgroups as moderators.

Risk of publication bias. Standard fail-safe N methods do not generalise to multiple dependent outcomes (e.g. Rosenthal, 1979; Orwin, 1983; Rosenberg, 2005). In such

conditions, sub-setting is required. Table 5 shows fail-safe Ns for Rosenthal, Orwin, and Rosenberg methods, grouped by control group and outcome measure combination. (See primary studies for details of the technical differences between methods.)

[Table 5]

The above estimates vary considerably, with a range of N=5 to N=2178 and substantial variation between group and outcome combinations across methods. Given a total cohort of nine studies, however, it may be reasonable to tentatively assume a low risk of publication bias significantly impacting the results reported above. Nevertheless, there does exist unobtainable data that may have shifted the estimates presented in Table 4: Two applicable studies were not included in the meta-analysis due to insufficient data to calculate standardised mean differences and sampling variances (Pizzioli & Schelstraete 2007, 2011), and an additional unpublished dataset was identified though not retrieved through expert consultation.

Discussion

This meta-analysis examined studies testing children with DLD on the auditory lexical decision task. Two thousand three hundred and seventy-two (2372) records were initially retrieved through electronic database searches, bibliographic searches, and expert consultation, with nine studies then selected for inclusion on the basis of eligibility and quality criteria. The final cohort included 499 children aged 3;8-11;4. The question examined was:

> *What are the estimated population effect sizes of the discrepancies in performance (response accuracy and latency) between children with DLD and age- and language-matched controls on the auditory lexical decision task?*

This question was addressed using a multivariate, random effects model. Estimates shown in Table 4 suggest children with DLD were considerably less accurate than age-matched

controls at identifying auditory words and rejecting auditory non-words, with a strong effect size estimate in this condition. However, the response time estimate for the same group comparison was less conclusive, with a confidence interval marginally crossing zero. This does not demonstrate no population effect, but indicates that zero or effect sizes approaching zero are a reasonable possibility for the underlying true effect. Thus while children with DLD appear considerably less accurate in the auditory lexical decision task than their age-matched peers, the current estimates suggest that they may not, in general, be significantly slower in making their responses. It is worth noting here that four primary studies investigated but found no evidence of a speed-accuracy trade-off, in which response accuracy may be compromised by a concern to provide a rapid response, or, alternatively, accuracy is high among participants who take considerable time planning a response (Crosbie et al., 2004; Edwards & Lahey, 1996; Pizzioli & Schelstraete, 2013; Quémart & Maillart, 2016). Note also that Edwards and Lahey (1996) and Crosbie et al. (2004) included a measure of auditory-vocal reaction time (AVRT), in which participants were required to say 'yes' immediately upon hearing a tone. In each study, analysis indicated no significant difference between experimental and control groups, suggesting between-group discrepancies in responses to lexical stimuli may not be attributable to difficulty identifying general forms of signal, formulating and articulating a verbal response, or the general complexity of task demands.

Confidence intervals for estimates of comparisons to language-matched controls crossed zero for both accuracy and response time outcomes, again suggesting zero is a reasonable possibility for the underlying true effect in these conditions. The observation that the accuracy deficit in particular 'disappears' when groups are matched by language ability is consistent with the view that the development of affected children is delayed though not deviant (Bishop, 1997), and provides tentative support for accounts specifying a causal

association between vocabulary growth and increasingly detailed lexical representations (e.g. Walley, 1993; Walley, Metsala, & Garlock, 2003). A 'delayed but not deviant' account of results should, however, be taken with some caution. As commented by an anonymous reviewer, the current study looks at just one type of task, and it is plausible that the same participants are delayed by different time intervals in different types of task; one year in task A, though two years in task B, etc. Indeed, below we discuss posited subgroups whose receptive vocabulary skills appear unimpaired despite receptive grammatical deficits warranting diagnosis. The term delayed may therefore be something of an oversimplification because the linguistic profile of a particular language-impaired child is unlikely to correspond to a discrete age range in typical development. The term delayed also suggests that these children will eventually catch up with peers, which is unclear from the data at hand.

In summary, results of the current meta-analysis are in line with previous reports of performance deficits among children with DLD relative to age-matched controls in tasks held to measure the accuracy of lexical representations (e.g. Borovsky, Burns, Elman, & Evans, 2013; Farquharson, Centanni, Franzluebbers, & Hogan, 2014; Ramus, Marshall, Rosen, & van der Lely, 2013; Rispens & Baker, 2012). However, this conclusion requires some qualification. The forest plots shown in Appendix B indicate variance in effect sizes both between and within studies, and the meta-analysis included a number of studies presenting results that contradict the overall estimates presented in Table 4: Befi-Lopes et al., (2010) and Maillart et al. (2004), for instance, report accuracy discrepancies between children with DLD and language-matched controls, while Edwards and Lahey (1996) and Pizzioli and Schelstraete (2013) report significant differences in response time between children with DLD and age-matched controls. In the sections that follow we discuss two broad factors that may contribute to such variability in outcomes: (a) sampling variation, within-group

heterogeneity, and the presence of possible sub-groups, and (b) study-specific differences in stimulus manipulation.

Sampling variation and sub-groups

Children with DLD differ widely in the specific problems they have with language. Unsurprisingly, then, primary studies included in the meta-analysis often reported relatively large variances among experimental groups (e.g. Crosbie et al., 2004), while in three studies subgroups associated with different performance profiles were formally identified (Crosbie et al., 2004; Edwards & Lahey, 1996; Windsor & Hwang, 1999). Such sub-group analyses are valuable because they may help explain how children with particular patterns of impairment approach the problem of spoken word recognition. Two studies sub-classified experimental-group participants on the basis of standardised test data (Edwards & Lahey, 1996; Windsor & Hwang, 1999). Edwards and Lahey (1996) report that children they describe as having expressive-only deficits performed considerably better than children with so-called mixed, or expressive-receptive deficits. This may be expected given that the auditory lexical decision task is itself a receptive measure. However, the authors note that their sub-group analysis is of questionable validity given a lack of appropriate statistical power, which post-hoc analysis estimated at just 26% (with a type I error rate of $\alpha = .05$ and a type II error rate of $\beta = .20$, power should be .80, or 80%). Analysis of the same sub-groups by Windsor and Hwang (1999) was statistically inconclusive, and results were omitted from the published manuscript. Given sample sizes comparable to those in Edwards and Lahey (1996), however, it is likely that Windsor and Hwang's (1999) analysis was similarly underpowered. On analysis of their auditory lexical decision task data, Crosbie et al. (2004) identify a sub-group of children described as having pronounced 'lexical' deficits, who performed worse than age-matched controls, and a 'post-lexical', syntactic or integration deficit sub-group who performed in line with age-matched controls. As Crosbie et al. (2004) note, however, the

post-hoc identification of sub-groups is unsatisfactory, particularly given this approaches close association with questionable research practices such as *p*-hacking (or data dredging), in which data is mined for significant patterns not included in pre-specified hypotheses. In summary, the few existing attempts to accommodate experimental group heterogeneity and sub-groups are insufficient, and this prevented against the inclusion of linguistic sub-group as a moderator in our statistical model. Further studies pre-registering sub-groups of interest and conducting prospective power analyses are required to determine the impact of linguistic sub-profile on auditory lexical decision task performance. That said, a number of researchers have emphasised the need to look at DLD in terms of dimensions of impairment rather than discrete subtypes (e.g. Bishop, 2006, p. 220). One useful direction, therefore, may be to use standardised assessment scores as continuous predictors of task performance in a linear regression model, rather than defining categorical subgroups (e.g. lexical versus post-lexical) for use in *t*-tests or ANOVAs.

Study-specific stimulus manipulation

There was broad consensus that the rejection of non-words was slower and less accurate across groups than responses to words, with this pattern typically pronounced in DLD groups relative to age-matched controls (Edwards & Lahey, 1996; Haebig et al., 2015; Pizzioli & Schelstraete, 2013). This finding may be attributable to the absence of long-term representations corresponding to non-words prompting extended lexical searches, though positive response bias may also play a role.

Word and non-word stimuli were often further manipulated in line with the primary study aims. Haebig et al. (2015), for instance, manipulated semantic network size to examine the role of semantics in spoken word processing by children with DLD and autism spectrum disorder. In this study, stimuli comprised twenty target words with a high number of semantically associated words, and twenty target words with a low number of semantically

associated words. There is no question that such manipulations contribute to variability in effect sizes. For instance, the discrepancy highlighted above between cases 87 and 88 from Haebig et al. (2015) may be attributable to the use of high- and low-semantic network words respectively. However, because such variables were often study specific, we considered their formal inclusion as moderators in our model to be of questionable value. Position and degree of non-word manipulation (see introduction for examples) were included as independent variables in two out of nine studies (Befi-Lopes et al., 2010; Maillart et al., 2004), which we again considered insufficient to warrant the inclusion of these variables as moderators. Interestingly, however, these studies showed considerable disagreement. Maillart et al. (2004), for instance, report that children with DLD were relatively less accurate when non-word manipulations occurred in initial or final (though not medial) positions, while Befi-Lopes et al. (2010) report a word initial manipulation advantage relative to language-matched controls for children with DLD lexical age 5;0. It is plausible that this disagreement is attributable in part to assessing samples with different first languages (French- and Brazilian-Portuguese-speaking respectively), with the word regions most amenable to segmentation apparently moderated by the phonological system of a particular language (van der Feest & Fikkert, 2015). This example illustrates the complex interaction between fine-grained differences in stimulus type and the sampling variation discussed above. In summary, lack of an appropriate number of studies incorporating comparable stimulus sub-classifications made it unclear what conclusions could be drawn had we included these variables as moderators, though identifying the specific characteristics that make a non-word difficult for children with DLD to accurately reject undoubtedly constitutes an important part of the future research agenda. Researchers interested in examining the variables discussed in this section in more detail may consult the associated R file, in which all sub-classifications are coded as part of the master dataset.

Limitations of the study cohort

This section addresses what we consider limitations of the study cohort, with the aim of improving future research using the auditory lexical decision task. First, we are aware of no paper including formal reliability and validity estimates with respect to the auditory lexical decision task. A useful model for this line of inquiry is a recent paper by West, Vadillo, Shanks, and Hulme (2017), who conducted reliability analyses into a range of tasks thought to measure procedural learning. These researchers report low task reliability and a prevalence of so-called 'extreme groups', which may overestimate the extent of linear relationships between variables in the population (p. 11). It is unclear whether the lexical underspecification literature suffers from similar issues, though a partial replication of West et al.'s (2017) study in this domain would be welcome considering the diversity of paradigms argued to converge on the quality of lexical representations (e.g. gating, naming, non-word repetition, and lexical decision), as well as widespread inconsistency in outcomes.

Second, a number of studies provided no conclusive evidence that the words used at test were known to the participant (e.g. Windsor & Hwang, 1999). In some research contexts this may be unnecessary. However, in auditory lexical decision studies claiming to assess the quality of lexical representations in long-term memory, it is essential to confirm that participants know the test words. Using normative data is common, though may not be appropriate in language-impaired samples unless carefully adjusted. Explicitly testing word knowledge prior to measuring auditory lexical decisions at delayed test may be preferable.

Third, designs showed variation in both the response required by children (e.g. verbal–'yes'/'no'–or non-verbal–ergonomic box or computer screen with green/red or smiley/sad-face buttons) and the method of auditory stimulus presentation (i.e. pre-recorded or spoken live by an experimenter; see Appendix A). While these dissimilarities appear trivial, they can introduce systematic bias. For instance, Maillart et al. (2004) describe a pilot

study in which children with DLD responded differently to pre-recorded stimuli presented via computer and stimuli spoken live by an experimenter, arguably due to adopting a compensatory strategy involving visual cues (i.e. lip reading) to aid target word discrimination (see also Bishop, Brown, & Robson, 1990). Such accounts reaffirm that researchers must carefully consider and justify each methodological decision.

Fourth, the focus of the current meta-analysis has been behavioural studies. However, McArthur and Bishop (2005) note that one limitation of the use of behavioural paradigms with children with DLD is that results below criterion could reflect low attention or motivation rather than linguistic deficits. Considering this, future research using the auditory lexical decision task may benefit from integrating neuroimaging methods assumed less susceptible to disruption by fluctuations in attentiveness, e.g. electroencephalography.

Limitations of the current review

This meta-analysis attempted to reduce error and bias by following PRISMA guidelines, applying explicit eligibility and quality criteria, pre-registering a research protocol, soliciting expert consultation, using multiple coders, and making the associated R code publicly available. Notwithstanding this methodological thoroughness, the current analysis has a number of limitations. First, three known datasets were omitted due to reporting insufficient statistics (n=2) or no longer being available (n=1; the latter was identified through expert consultation). Importantly, two of these studies (Pizzioli & Schelstraete 2007, 2011), reported no significant difference in response accuracy between children with DLD and age-matched controls, and so contradict the strong effect size estimate reported in Table 4. Unfortunately, the degree to which the inclusion of these datasets would have affected the population estimates presented in the current meta-analysis is unclear. Second, heterogeneity in the population of children with DLD along with unsatisfactory attempts to accommodate posited subgroups in the primary literature restrict

the extent to which we are currently able to generalise findings to the population. Relatedly, it is regrettable that the numbers of primary studies incorporating particular stimulus sub-classifications were not sufficient to warrant the inclusion of these variables as moderators.

Conclusion

Despite the apparent primacy of syntactic deficits, children with DLD often evidence lexical impairments. In particular, it has been argued that this population has difficulty forming lexical representations detailed enough to enable them to process spoken words efficiently. The current meta-analysis examined studies testing children with DLD in the auditory lexical decision task; a behavioural paradigm commonly used in both clinical and non-clinical research contexts to assess the quality of lexical representations. Effect size estimates suggest children with DLD were less accurate though not necessarily slower in this task than age-matched controls, with no significant difference with respect to either accuracy or response time between children with DLD and language-matched controls. The primary literature provides suggestive evidence that the observed accuracy deficit may not be attributable to a speed accuracy trade-off, difficulty identifying general forms of signal, formulating and articulating a verbal response, or the complexity of task demands. Future research using the auditory lexical decision task should address the issue of within-group heterogeneity by pre-registering experimental sub-groups of possible interest or using continuous rather than categorical predictors, and conducting prospective power analyses to determine adequate sample sizes. Better understanding of the specific lexical characteristics (e.g. the position of non-word manipulation) that make an auditory non-word difficult for certain children with DLD to reject is also required. Finally, reliability analysis constitutes an important part of the future research agenda given inconsistencies in the existing literature on lexical underspecification.

Acknowledgements

Thank you to Barbara Dodd, Shelley Gray, Eileen Haebig, Suze Leitão, Franck Ramus, Phaedra Royle, Richard Schwartz, Bill Wells, Cori Williams, and Jennifer Windsor for providing feedback, research papers, and data in response to consultation requests. Thank you also to Daniel Quintana, Andy Field, and Wolfgang Viechtbauer, for their comments on methodology.

References

Ainsworth, S., Welbourne, S., & Hesketh, A. (2016). Lexical restructuring in preliterate children: Evidence from novel measures of phonological representation. *Applied Psycholinguistics*, *37*(4), 997–1023. https://doi.org/10.1017/S0142716415000338

Almodovar, D. (2014). *Effects of phonological neighborhood density on lexical access in adults and children with and without specific language impairment*. *City University of New York (CUNY) Academic Works*. The City University of New York. Retrieved from http://academicworks.cuny.edu/gc_etds/160/?utm_source=academicworks.cuny.edu%2Fgc_etds%2F160&utm_medium=PDF&utm_campaign=PDFCoverPages

Befi-Lopes, D. M., Pereira, A. C. S., & Bento, A. C. P. (2010). Phonological representation of children with specific language impairment (SLI). *Pró-Fono*, *22*(3), 305–310. https://doi.org/S0104-56872010000300025

Bishop, D. V. M. (1997). *Uncommon understanding: Development and disorders of language comprehension in children.* Hove, England, UK: Taylor & Francis.

Bishop, D. V. M. (2006). What causes specific language impairment in children? *Current Directions in Psychological Science*, *15*(5), 217–221. https://doi.org/10.1111/j.1467-8721.2006.00439.x

Bishop, D. V. M., Brown, B. B., & Robson, J. (1990). The relationship between phoneme discrimination, speech production, and language comprehension in cerebral-palsied individuals. *Journal of Speech and Hearing Research*, *33*, 210–219. https://doi.org/10.1044/jshr.3302.210

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., & Greenhalgh, T. (2016). CATALISE: A multinational and multidisciplinary delphi consensus study. Identifying language impairments in children. *PLOS ONE*, *11*(7), e0158753. https://doi.org/10.1371/journal.pone.0158753

Borovsky, A., Burns, E., Elman, J. L., & Evans, J. L. (2013). Lexical activation during sentence comprehension in adolescents with history of specific language impairment. *Journal of Communication Disorders*, *46*, 413–427. https://doi.org/10.1016/j.jcomdis.2013.09.001

Claessen, M., & Leitão, S. (2012). Phonological representations in children with SLI. *Child Language Teaching and Therapy*, *28*(2), 211–223. https://doi.org/10.1177/0265659012436851

Crosbie, S. L., Howard, D., & Dodd, B. J. (2004). Auditory lexical decisions in children with specific language impairment. *British Journal of Developmental Psychology*, *22*(1), 103–121. https://doi.org/10.1348/026151004772901131

Dollaghan, C. (1998). Spoken word recognition in children with and without specific language impairment. *Applied Psycholinguistics*, *19*(2), 193–207. https://doi.org/10.1017/S0142716400010031

Edwards, J., & Lahey, M. (1996). Auditory lexical decision of children with specific language impairment. *Journal of Speech and Hearing Research*, *39*, 1263–1273.

Farquharson, K., Centanni, T. M., Franzluebbers, C. E., & Hogan, T. P. (2014). Phonological and lexical influences on phonological awareness in children with specific language impairment and dyslexia. *Frontiers in Psychology*, *5*, 1–10. https://doi.org/10.3389/fpsyg.2014.00838

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*

(2nd ed.) (pp. 357–376). New York: Russell Sage Foundation.

Goodman, E., & Bates, J. C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processes*, *12*(5–6), 507–584. https://doi.org/10.1080/016909697386628

Graf Estes, K., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *50*(1), 177–195. https://doi.org/10.1044/1092-4388(2007/015)

Haebig, E., Kaushanskaya, M., & Ellis Weismer, S. (2015). Lexical processing in school-age children with autism spectrum disorder and children with specific language impairment: The role of semantics. *Journal of Autism and Developmental Disorders*, *45*(12), 4109–4123. https://doi.org/10.1007/s10803-015-2534-2

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

James, D., Van Steenbrugge, W., & Chiveralls, K. (1994). Underlying deficits in language-disordered children with central auditory processing difficulties. *Applied Psycholinguistics*, *15*(3), 311–328. https://doi.org/10.1017/S0142716400065917

Kan, P. F., & Windsor, J. (2010). Word Learning in Children With Primary Language Impairment: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *53*(3), 739–756. https://doi.org/10.1044/1092-4388(2009/08-0248)

Maillart, C., Schelstraete, M.-A., & Hupet, M. (2004). Phonological representations in children with SLI: A study of French. *Journal of Speech, Language, and Hearing Research*, *47*(1), 187–198. https://doi.org/10.1044/1092-4388(2004/016)

McArthur, G. M., & Bishop, D. V. M. (2005). Speech and non-speech processing in people

with specific language impairment: A behavioural and electrophysiological study. *Brain and Language*, *94*(3), 260–273. https://doi.org/10.1016/j.bandl.2005.01.002

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. London: Taylor and Francis. https://doi.org/10.3102/1076998616659752

Montgomery, J. W. (1999). Recognition of gated words by children with specific language impairment: An examination of lexical mapping. *Journal of Speech, Language, and Hearing Research*, *43*(3), 735–743.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*. https://doi.org/10.2307/1164923

Pizzioli, F., & Schelstraete, M.-A. (2007). *Auditory lexical decision in children with specific language impairment*. *Proceedings of the 31st Boston University Conference on Language Development*. https://doi.org/10.1348/026151004772901131

Pizzioli, F., & Schelstraete, M.-A. (2011). Lexico-semantic processing in children with specific language impairment: The overactivation hypothesis. *Journal of Communication Disorders*, *44*(1), 75–90. https://doi.org/10.1016/j.jcomdis.2010.07.004

Pizzioli, F., & Schelstraete, M.-A. (2013). Real-time sentence processing in children with specific language impairment: The contribution of lexicosemantic, syntactic, and world-knowledge information. *Applied Psycholinguistics*, *34*, 1–30. https://doi.org/10.1017/S014271641100066X

Quémart, P., & Maillart, C. (2016). The sensitivity of children with SLI to phonotactic probabilities during lexical access. *Journal of Communication Disorders*, *61*, 48–59. https://doi.org/10.1016/j.jcomdis.2016.03.005

Ramus, F., Marshall, C. R., Rosen, S., & van der Lely, H. K. J. (2013). Phonological deficits in specific language impairment and developmental dyslexia: Towards a multidimensional model. *Brain*, *136*(2), 630–645. https://doi.org/10.1093/brain/aws356

Rispens, J., & Baker, A. (2012). Nonword repetition: The relative contributions of phonological short-term memory and phonological representations in children with language and reading impairment. *Journal of Speech, Language, and Hearing Research*, *55*(3), 683. https://doi.org/10.1044/1092-4388(2011/10-0263)

Robey, R. R., & Dalebout, S. D. (1998). A tutorial on conducting meta-analyses of clinical outcome research. *Journal of Speech, Language, and Hearing Research*, *41*, 1227–1241. https://doi.org/1092-4388/98/4106-1227

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*(2), 464–468. https://doi.org/10.1111/j.1095-8649.2006.01157.x

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

van der Feest, S. V. H., & Fikkert, P. (2015). Building phonological lexical representations. *Phonology*, *32*(2), 207–239. https://doi.org/10.1017/S0952675715000135

Viechtbauer, W. (2010). Conducting meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Walley, A. C. (1993). The role of vocabulary development in children′s spoken word recognition and segmentation ability. *Developmental Review*, *13*(3), 286–350. https://doi.org/10.1006/drev.1993.1015

West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2017). The procedural learning deficit hypothesis of language learning disorders: we see some problems. *Developmental Science*, (May 2016), 1–13. https://doi.org/10.1111/desc.12552

Windsor, J., & Hwang, M. (1999). Children's auditory lexical decisions: A limited processing capacity account of language impairment. *Journal of Speech, Language, and Hearing Research*, *42*(4), 990–1002. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/10912254