Journal of Memory and Language 141 (2025) 104586

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

The separability of early vocabulary and grammar knowledge

Seamus Donnelly [a,b,*], Evan Kidd [b,c], Jay Verkuilen [d], Caroline Rowland [b,e,f]

[a] School of Medicine and Psychology, The Australian National University, Australia
[b] Language Development Department, Max Planck Institute for Psycholinguistics, the Netherlands
[c] School of Languages, Literature and Linguistics, The Australian National University, Australia
[d] Program in Educational Psychology, The Graduate Center, City University of New York, USA
[e] Donders Centre for Brain, Language and Cognition, Radboud University, the Netherlands
[f] Department of Psychology, University of Liverpool, UK

ARTICLE INFO

Keywords:
Language development
Grammar
Vocabulary
Psychometrics

ABSTRACT

A long-standing question in language development concerns the nature of the relationship between early lexical and grammatical knowledge. The very strong correlation between the two has led some to argue that lexical and grammatical knowledge may be inseparable, consistent with psycholinguistic theories that eschew a distinction between the two systems. However, little research has explicitly examined whether early lexical and grammatical knowledge are statistically separable. Moreover, there are two under-appreciated methodological challenges in such research. First, the relationship between lexical and grammatical knowledge may change during development. Second, non-linear mappings between true and observed scores on scales of lexical and grammatical knowledge could lead to spurious multidimensionality. In the present study, we overcome these challenges by using vocabulary and grammar data from several developmental time points and a statistical method robust to such non-linear mappings. In Study 1, we examined item-level vocabulary and grammar data from two American English samples from a large online repository of data from studies employing a commonly used language development scale. We found clear evidence that vocabulary and grammar were separable by two years of age. In Study 2, we combined data from two longitudinal studies of language acquisition that used the same scale (at 18/19, 21, 24 and 30 months) and found evidence that vocabulary and grammar were, under some conditions, separable by 18 months. Results indicate that, while there is clearly a very strong relationship between vocabulary and grammar knowledge in early language development, the two are separable. Implications for the mechanisms underlying language development are discussed.

Introduction

Children face two intertwined problems during language acquisition: learning words and the grammatical processes that operate over them. Any theory of language acquisition must account for how children learn both. Historically, theories of language development have fallen in two camps: (i) dual-systems approaches (e.g., 'words and rules', Pinker, 1999), which assume these are separate problems underwritten by separate learning systems, and (ii) lexicalist approaches, which assume that the same learning mechanism supports acquisition in both domains, such that grammatical knowledge is a consequence of lexical learning (Bates & Goodman, 1997; Plunkett & Marchman, 1993). It has proven difficult to adjudicate between these accounts as their specific predictions are determined by often unarticulated implementational assumptions about the format of linguistic representations and the learning mechanisms that support development. As a result, the current theoretical landscape is vast and varied, complicating attempts at empirical work that aims to directly compare competing theories. However, nearly all existing theories make assumptions about the separability of grammatical and lexical knowledge, particularly in the earliest years of language acquisition. Clear evidence for the separability (or lack thereof) of lexical and grammatical knowledge, particularly at the earliest stages of language development, would establish important boundaries between the classes of plausible theoretical accounts.

While the relationship between early lexical and grammatical knowledge is well studied, less work has examined the dimensional structure of early vocabulary and grammatical knowledge, and the available research is difficult to interpret in light of underappreciated

* Corresponding author at: Seamus Donnelly, School of Medicine and Psychology, The Australian National University, Canberra 2601, Australia.
E-mail address: seamus.donnelly@anu.edu.au (S. Donnelly).

https://doi.org/10.1016/j.jml.2024.104586
Received 24 August 2023; Received in revised form 22 November 2024; Accepted 24 November 2024
Available online 29 December 2024

methodological and statistical challenges. First, interactions between initially separate vocabulary and grammar subsystems could cause the two systems to become statistically inseparable, such that what looks like a single system is built upon initially differentiable competencies (Van der Maas et al. 2006). Second, non-linear mappings between true and observed measures could cause a truly one-dimensional system to appear multidimensional (Dixon & Marchman, 2007; Hayes et al., 2017; Stephens et al., 2019). The present work aims to overcome these limitations by testing the dimensionality of vocabulary and grammatical knowledge at multiple time points using a statistical method robust to potential non-linear mappings. In doing so, we provide the most robust, comprehensive examination of the developmental relationship between vocabulary and grammar in early child development to date, and provide an important developmental benchmark for a question at the heart of core theoretical debates in language acquisition

The inseparability of lexicon and grammar?

Many theories in linguistics and psychology have assumed that grammar and vocabulary reflect distinct representations and processes (Pinker, 1999), which we call *dual-systems approaches.* While some accounts assume these are learned by separate systems from day 1 (Gertner et al., 2006; Pinker, 1999) and others assume children learn initially item-specific representations over which they abstract to eventually construct grammatical generalizations (Bannard et al., 2009), these theories share the assumption that children eventually produce strings of words structured by abstract grammatical knowledge that is not fully reducible to lexical representations. An alternative account is that lexical and grammatical knowledge are learned and processed through the same sets of cognitive processes (Ambridge, 2020a; Bates & Goodman, 1997; Fedorenko et al., 2020), a view inspired by work in neural-network based models (Bates & Goodman, 1997; Plunkett & Marchman, 1993), and reflected in exemplar-based models (Ambridge, 2020a), which assume that lexical and grammatical knowledge are stored and processed in the same way. Such accounts are consistent with recent neuroimaging studies in adult populations that observe no cortical selectivity for lexical vs grammatical tasks (Fedorenko et al., 2016, 2020; Shain et al., 2020). Given the assumption on these views that grammatical knowledge is an emergent property of the lexicon, we call these single-system approaches *lexicalist approaches*.

One of the most striking pieces of evidence for lexicalist approaches to acquisition is the impressively strong correlation between vocabulary and grammar measures in early child language development (Bates & Goodman, 1997; Brinchmann et al., 2019; Dale et al., 2000; Dionne et al., 2003; Frank et al., 2021; Hoff et al., 2018; Valentini & Serratrice, 2021). For example, in a seminal article, Bates and Goodman (1997) report that the correlation between vocabulary and grammar on the MacArthur Bates Communicative Development: Words and Sentences (MBCDI; Marchman et al., 2023) was very strong ($r = .84$). Moreover, this relationship was stronger than the relationship between productive and receptive vocabulary. They also noted that cross-lagged correlations between vocabulary and mean length of utterance (MLU), a common index of morphosyntactic ability, often approached or equaled the size of cross-lagged correlations within both domains. Bates and Goodman (1997) argued that these findings reflected the *inseparability* of lexical and grammatical knowledge, consistent with the lexicalist approaches described above.

While many studies have observed strong correlations between lexical and vocabulary development (Bates & Goodman, 1997; Brinchmann et al., 2019; Dale et al., 2000; Dionne et al., 2003; Frank et al., 2021; Hoff et al., 2018; Valentini & Serratrice, 2021), very few have directly tested whether these correlations are better explained by models that assume one or two distinct systems, and existing studies yield contradictory results, perhaps reflecting the very different age groups tested. Two studies with older (school aged) children have suggested that vocabulary and grammatical knowledge are unidimensional at the onset of schooling (Language and Reading Research Consortium, 2015; Tomblin & Zhang, 2006). Tomblin and Zhang (2006) followed a cohort of children from kindergarten to grade 8 and administered batteries of grammatical and vocabulary knowledge at each time point. They found little evidence of separate dimensions for vocabulary and grammar at kindergarten (average age 6.04 years), second (7.47 years) and fourth grade (9.44 years), but clear evidence at eighth grade (average age 13.4 years). The Language and Reading Research Consortium (2015) observed a similar pattern of results with a different sample and different battery of tests. They administered a battery of grammar, vocabulary and discourse tasks to 955 children between prekindergarten (5 years) and Grade 3 (8.5 years) and did not find clear evidence of separate dimensions for vocabulary and grammar until Grade 3. In discussing their results together with those of Tomblin and Zhang (2006), the authors argued that the most parsimonious explanation is that vocabulary and grammar are initially unidimensional, but that they become increasingly separable as children become exposed to increasingly complex literary language in reading.

Results of studies with young, pre-literate, children are more mixed. In a classic study, Bates et al. (1988), found little evidence for a distinction between vocabulary and grammar in a group of children tested at 13 and 28 months, though they did observe evidence for distinctions between comprehension, production of rote forms and production of analyzed forms. A limitation of this study was that the sample was far smaller than what is typical for modern studies of latent variables ($N = 27$). In a more recent, better powered study with young children ($N = 2250$), Day and Elison (2022) found evidence for (at least) two dimensions much earlier than the onset of schooling (ages: 13 to 30 months). The goal of the study was to test whether closed-class vocabulary items, such as prepositions and question words, loaded on a lexical or grammatical factor, using data from the MBCDI from Wordbank (Frank et al., 2017), an open repository of data collected with various versions of the MBCDI and its various adaptations. The authors submitted the section subscores (Words about Time, Animals, Grammatical Complexity, etc) of the MBCDI Words and Sentences to exploratory and confirmatory factor analyses and observed that open class words loaded onto one factor but that the various grammatical subsections and vocabulary sections for closed-class items (e.g., Question Words) loaded on a second factor. This suggests that the two systems are separable from much earlier: from 2 years of age. One additional study (Pérez-Leroux et al., 2012) compared a single and multi-factor models of vocabulary and grammatical knowledge in a sample of 110 Spanish speaking children aged 3–5 and found that, while the one-factor model fit poorly, the four-factor model fit better. However, these results are difficult to interpret as the four-factor model included factors for vocabulary and grammar but also one factor for two blocks of a task measuring children's use of determiners and one for two blocks of a task measuring children's use of direct object pronouns. As a result, it's unclear if the superior performance of the four-factor model reflects a dissociation between vocabulary and grammar or the relatively low correlations between the tasks measuring performance on determiners and direct objects and the other tasks. Indeed, the vocabulary factor significantly predicted the grammar factor and the direct object factor, but the grammar and direct object factor were unrelated.

The inconsistent results across studies and samples may reflect two often unacknowledged challenges in research of this nature. First, if vocabulary and grammar exhibit mutual causal relationships, such that lexical and grammatical knowledge are initially separable, they could become so correlated over developmental time as to become inseparable (Van Der Maas et al., 2006). Thus, it is important to assess their relationship over time, with a particular focus on the earliest stages of language development. Second, if the relationships between measured and true scores for instruments used in these studies are non-linear, this could create spurious multidimensionality, even if the system were unidimensional (Hayes et al., 2017; Stephens et al., 2019). Thus, it is important to take account of possible non-linearity, or other possible

model misspecifications. We discuss each of these challenges in turn.

Developmental changes in the dimensional structure

A one-dimensional structure in older children could be a developmental outcome of two initially separate but mutually and causally related systems. Simulation research has shown that initially separate cognitive processes can become so correlated as to be statistically indistinguishable if there are reciprocal causal relationships between the systems (Van Der Maas et al., 2006), sometimes called *mutualistic coupling*. For example, in a set of simulation studies aimed at explaining the one-factor structure of intelligence tests, Van der Maas et al (2006) showed that a set of initially uncorrelated cognitive processes (e.g., working memory, analogical reasoning), which exhibited reciprocal causal relations (e.g., growth in working memory leads to growth in analogical reasoning and vice versa), became so correlated over development that they quickly exhibited the one-factor *g* structure of intelligence tests. Consistent with this, recent empirical work has shown that developmental changes in the reasoning and vocabulary subscales of a common IQ test are better explained by mutualistic coupling between the two systems than growth in a single underlying *g*-factor (Kievit et al., 2017; Kievit et al., 2019).

A similar dynamic could explain a one-dimensional structure found in vocabulary and grammar as there are several hypothesized pathways by which lexical knowledge could influence grammatical knowledge and vice versa. For example, knowledge of vocabulary could provide top-down support for the real-time perception and acquisition of morphosyntactic features (so-called *perceptual bootstrapping*, Nusbaum & Goodman (1994)), and knowledge of syntactic structure could aid the acquisition of verbs and other predicating words (so-called *syntactic bootstrapping*, Gleitman, 1990). Therefore, if vocabulary and grammar are initially separate systems, but increased proficiency in one facilitates learning in the other, they could become statistically unidimensional with time.

At first blush, this possibility seems inconsistent with structural equation modelling studies which typically fail to observe direct bidirectional relationships between lexical and grammatical knowledge (Brinchmann et al., 2019; Frank et al., 2021; Hoff et al., 2018; Valentini & Serratrice, 2021). However, conclusions from such studies are quite mixed, with some finding no direct relationships between grammatical and lexical development (Hoff et al., 2018; Valentini & Serratrice, 2021), others finding a direct effect of vocabulary on grammar between 16 and 30 months (Frank et al., 2021), and others finding a direct effect of grammar on vocabulary between 4 and 5 years (Brinchmann et al., 2019). Thus, while any individual study appears inconsistent with the mutualistic coupling, the totality of the evidence is less clear. Moreover, there are conceptual and methodological limitations with these cross-lagged studies which make evaluating mutualistic coupling difficult. First, these studies examine the effects of total vocabulary size and total grammar on one another. However, many of the proposed causal pathways between lexical and grammatical knowledge typically invoke more local causal relationships (e.g., syntactic bootstrapping means learning a particular syntactic structure will facilitate learning of a particular set of verbs that occur in that structure). Second, these studies have all relied on path analysis or structural equation modeling, which assume that true vocabulary and grammar relate linearly to measured vocabulary and measured grammar. This is unlikely to be true, especially with measures like the commonly used MBCDI and its adaptations (Dixon & Marchman, 2007),[1] a point to which we turn next.

Non-linear mappings between constructs and measures

A second challenge in statistically separating vocabulary and grammatical knowledge is the unknown relationship between true and measured scores on early language development instruments. Mental constructs such as vocabulary and grammatical knowledge are not directly observable. As a result, researchers rely on observable behavioural measurements, which are assumed to covary with the unobservable construct, to draw inferences about the unobserved construct. An underappreciated challenge in drawing these inferences is that the unobserved construct may relate non-linearly to the observed measure (Dixon & Marchman, 2007; Hayes et al., 2017; Stephens et al., 2019; Wagenmakers et al., 2012). Consider, for example, the Vocabulary subscale of a MBCDI. Given that the 680 items on the vocabulary subscale does not represent an exhaustive list of all words a young child might know, it is almost certainly non-linearly related to their true, underlying total vocabulary (Dixon & Marchman, 2007; Mayor & Plunkett, 2011).

In particular, consider Fig. 1A, which depicts a linear mapping between measured scores on the CDI and true total vocabulary for four hypothetical children, each represented by a different colour line. The parents of Child X (the blue line), Child Y (the red line), Child Z (the green line) and Child Q (the yellow line) checked off 0, 1, 501, and 502 words respectively on the Vocabulary subscale of the CDI. Because the relationship between measured and true vocabulary is linear, Children X and Y and Children Z and Q both differ by the same amount in their true vocabularies (in this case, 1 word). However, this scenario is unlikely to reflect the relationship between measured and true vocabulary in real children. Specifically, it is very unlikely to be the case that the Vocabulary subscore is linearly related to true vocabulary, such that every one-unit difference on the scale of the Vocabulary subscore corresponds to a constant difference on the scale of true vocabulary. Fig. 1B depicts an alternative, more plausible non-linear, relationship where Children Z and Q differ by more in true vocabulary than Children X and Y do, despite differing by the same amount on measured vocabulary. A similar non-linear mapping is likely to exist for the grammar section of CDIs as well.

There are several conceptual and empirical reasons to believe that the relationships between true and measured vocabulary, and true and measured grammar, on the MBCDI and related instruments are non-linear (i.e. that Fig. 1B, or some other non-linear function, reflects a more plausible scenario than Fig. 1A). First, both the Vocabulary and Grammar subscales have ceilings, but true vocabulary and grammar likely do not have ceilings that are relevant when studying young children, which would ensure some non-linearity in the relationship between the two. Second, Dixon and Marchman (2007) have shown that the well-known non-linear relationship between measured vocabulary and measured grammar could have been an artifact of the non-linear mapping between true and measured scores outlined above. To test this, they regressed Vocabulary and Grammar on one another and examined the residuals for evidence of non-linear mappings. They found that their data could be explained by assuming that true and measured grammar were non-linearly related. Third, Mayor and Plunkett (2011) tried to derive a total-score vocabulary estimate from CDI scores and found a strongly non-linear relationship, with initial increments in the CDI corresponding to smaller increments in total vocabulary than later increments. Finally, two recent studies controlling for Vocabulary subscores at time $t-1$, when modeling Vocabulary subscores at time t, found the relationship to be non-linear (Creaghe et al., 2021; Donnelly & Kidd, 2020), which would be expected if the relationship between true and observed scores was different across samples with different means. Together, these findings suggest a non-linear mapping between true and measured scores. We note that this possibility is not specific to CDI measures and that similar arguments have been made about experimental measures such as reaction times and accuracy (Loftus, 1978; Stephens et al., 2019; Wagenmakers et al., 2012).

[1] Throughout this paper we use MBCDI for the MacArthur Bates Communicative Development Inventory and CDI for its adaptations and to refer to the broader class of instruments.

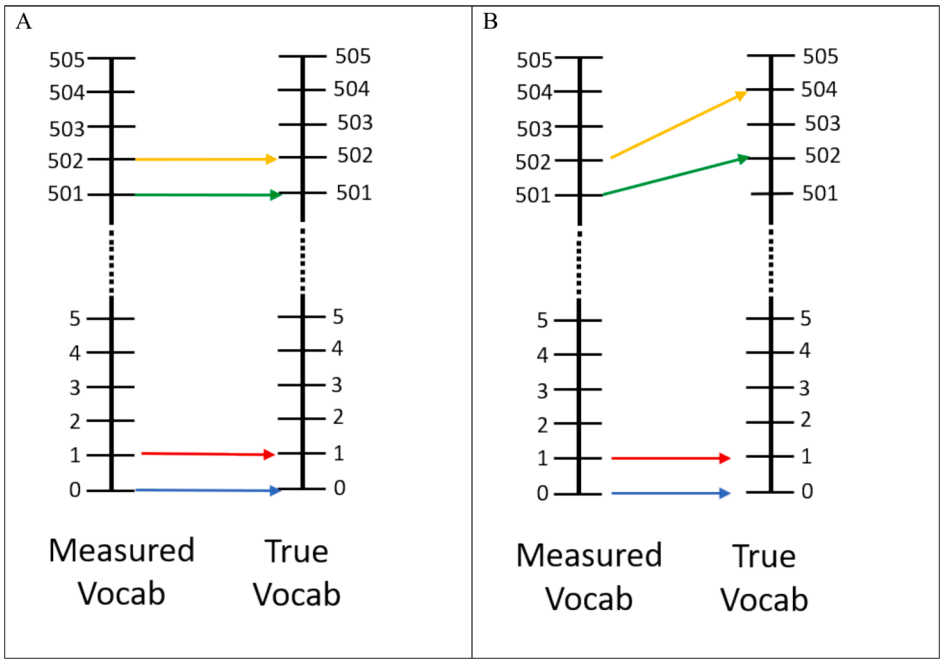

Fig. 1. Depiction of linear (A) and non-linear (B) mapping between true and measured vocabulary. In each pane, the scale on the lefthand side represents measured vocabulary and the scale one the righthand side represents true vocabulary. Each child is represented by a different colour line. As can be seen in Pane A, if the relationship between true and measured scores is linear, two pairs of children who differ by the same amount on measured vocabulary should differ by the same amount of true vocabulary. For example, the yellow and green children differ by 1 point on measured vocabulary and the red and blue children also differ by 1 point on measured vocabulary. Both pairs of children also differ by the same amount of true vocabulary (in this case, 1 point). As can be seen in Pane B, if the relationship between true and measured scores is non-linear two pairs of children who differ by the same amount on measured vocabulary may differ by different amounts on true vocabulary. For example, while both the yellow and green children and red and blue children differ by one point on measured vocabulary, the pairs differ by different amounts on true vocabulary. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

While this non-linear mapping does not pose a challenge in testing hypotheses which make ordinal predictions between groups of participants (for example that $M_{exp} > M_{control}$), it poses a significant challenge for individual differences research, which seeks to quantify variability between individuals, often on a metric scale (Kidd et al., 2018; Kidd & Donnelly, 2020). For example, given the findings of Mayor and Plunkett (2011), the difference between Children Z and Q in the scenario above is probably substantially larger than the difference between Children X and Y, but traditional statistical methods applied to raw CDI data assume these differences are identical. Failing to account for the non-linear mapping between true and measured variables can lead researchers to incorrect conclusions about the dimensionality of cognitive processes, which has led to warnings about interpreting non-crossover (sometimes called removeable) interactions (Loftus, 1978; Wagenmakers et al., 2012). Recently, Stephens et al (2019) showed that interaction effects often interpreted as evidence for two-systems accounts of category learning and decision making are equally consistent with one-systems accounts, if a non-linear mapping is assumed between the measured and observed variables. Likewise, Hayes et al (2017) found that developmental dissociations often assumed to reflect two distinct memory systems (recognition and recall) could also be explained by one memory system with non-linear mappings between observed and measured variables. If this potential non-linear mapping is not considered, a truly unidimensional system may appear to be multidimensional because of these mappings. One potential solution is to try to identify the function mapping true to observed scores and include that as a link function; however, mis-specification of this function could result in the same misleading pattern of multi-dimensionality. An alternative approach is to consider non-parametric or semi-parametric methods of dimensionality assessment, which make limited assumptions about the data-generating process.

Statistical approaches to dimensionality assessment

While the role of model misspecifications, such as incorrectly assuming linear relationships, in creating spurious multidimensionality is often overlooked in cognitive psychology [though see Dixon & Marchman (2007), Stephens et al. (2019) and Wagenmakers et al. (2012)], it is well studied in educational measurement and testing (Reckase, 2009). As large-scale educational tests are often designed to measure multiple distinct but overlapping domains, testing whether two sets of items on a single test– for examplealgebra and geometry items– reflect two distinct, albeit correlated, domains of between-individual variability is a key question in educational testing. Items on such tests are often analysed using the item-response theory framework (IRT), which is incorporated in the present work and increasingly used to analyze data from child language inventories such as CDIs (Frank et al., 2021).

IRT is a family of latent variable models for modeling item-level categorical data. Because IRT models make specific parametric assumptions about how true scores map on to observed scores, they can model specific non-linear relationships. For example, one of the most common models, the two-parameter logistic model (2PL), assumes that a participant *i*'s performance on a given item *j* is jointly determined by (i) the participant's latent ability (θ), (ii) the item's difficulty (*d*), and (iii) the item's discrimination parameter (*a*), which allows the item response functions to be non-parallel for different items.

$$y_{i,j} = a_j{}^*(\theta_i - d_j) \quad (1)$$

As the product of these terms can go above 1 and below 0, this product is mapped to the probability space via the logistic link function, as in 2:

$$p_{i,j} = logistic(a_j{}^*(\theta_i - d_j)) \quad (2)$$

Because of the logistic link function, this model assumes that the latent participant ability maps to observed scores following a logistic function, thereby explicitly modeling a particular type of non-linear relationship between true and observed item scores. It is possible to extend this model by adding a second ability parameter (θ_2). Doing so would allow us to estimate one and two-factor models, as in confirmatory factor analysis.

Unfortunately, while the 2PL model accommodates a form of non-linearity, it makes two specific parametric assumptions that may be incorrect. First, it assumes that the latent participant abilities are normally distributed, ranging between negative and positive infinity. Second, it assumes that the relationship between the latent proficiencies and observed scores follows a logistic link function. As violations of these assumptions could create artifactual multidimensionality, just like violation of linearity in traditional regression and factor analysis approaches, their validity is critical. The first assumption is unlikely to hold, given that vocabulary and grammar certainly have floors on the construct level, and the second assumption is difficult to evaluate. Huang and Bolt (2024) argue that both of these assumptions are unlikely to hold in many empirical domains in psychology, including vocabulary, and advocate for the use of theoretically informed models that make task-specific assumptions about both. However, they also demonstrate that these assumptions are difficult to test empirically; in particular, they show that while the 2PL model and a more theory-informed, task-specific model provide nearly identical fits to the same dataset, the two models make different predictions about the distribution of true scores and the link between true and observed scores.

An alternative approach to making specific parametric assumptions about the data-generating process is using non-parametric item-response theory models, which make much more limited assumptions than parametric models, such as the 2PL model, do. These assumptions are known as *weak*, which is to say that they use first- and second-order quantities such as means, variances, and covariances as opposed to specific link functions or likelihoods, which are known as *strong* assumptions. In a comparison of parametric and non-parametric IRT models, Sijtsma and Meijer (2006) find that nonparametric IRT is generally better at tasks such as anomaly detection and dimensionality assessment, while parametric IRT is generally more efficient statistically and better able to deal with incompletely observed data. This mirrors the difference between parametric and non-parametric methods in experimental psychology, where parametric methods like regression and analysis of variance are more powerful and flexible because of the constraints imposed by their parametric assumptions but can produce artifactual results if those assumptions are strongly violated. In contrast, as non-parametric methods make more general assumptions, they are compatible with a wider range of observed data sets but considerably less flexible in application. This is one reason by non-parametric IRT models are viewed as attractive tools dimensionality assessment (e.g., Bonifay et al., 2015; Sijtsma and Meijer, 2006; Stout et al., 1996; Zhang & Stout 1999; Jang & Roussos, 2007). Nonparametric approaches can also cope with smaller sample sizes due to the fact that a complex model is not being estimated.

One particularly attractive non-parametric method, DETECT (*Dimensionality Evaluation to Enumerate Contributing Traits*, Stout et al., 1996), relies on three key assumptions: monotone increasing link functions, continuous latent variables, and local independence, which are the minimal assumptions needed for an IRT model. In other words, DETECT makes no specific assumption about the shape of the latent variable distribution and merely assumes that the link function is monotone increasing. To determine whether multidimensionality is present in the data, DETECT makes use of a decomposition of item variation into the dominant dimension and dimensions orthogonal to it. In particular, it examines residual item pair covariation after removing the dominant dimension, by conditioning on a composite score calculated using all items (using, for example, participant ability scores from the 2PL model). If items *i* and *j* measure the same dimension, then, upon partialling out shared variability in the composite score, item pair covariance should be positive. If these items measure separate dimensions, then, upon partialling, their covariance should be negative. Confirmatory DETECT averages the conditional covariances hypothesized to reflect the same cluster and −1 times the conditional covariances between items hypothesized to reflect different clusters. The resulting DETECT index indicates the degree of departure from unidimensionality and can be understood as an effect size (Jang & Roussos, 2007).

DETECT was developed primarily to test dimensionality in educational testing data and has been used to evaluate the dimensional structure of the Test of English as a Foreign Language (TOEFL; Jang & Roussos, 2007), the Literacy Assessment and Monitoring Program (Zhang, 2013), the Law School Admissions Test (Kim, 1994). While much of the published work on DETECT is methodological in nature, Jang and Roussos (2007) used DETECT to analyze the factor structure of various sections of the TOEFL and determined that the scale was best described with two factors, one for listening comprehension and one for reading comprehension and structure/written expression (SWE). In particular, they reported DETECT indices between .31 and .37 to show that the reading comprehension and SWE sections of TOEFL reflect a different dimension than the listening comprehension section, concluding that the listening comprehension is a construct that reflects a "significant and substantial difference from the ability to understand written English" (Jang & Roussos, 2007, p. 10). When they analyzed items from just the reading comprehension and SWE sections, they observed DETECT indices between .15 and .17 and concluded these measures reflected the same underlying construct.

While DETECT has not been widely used outside of educational testing, it offers several attractive features for testing the dimensionality of CDIs. First, the primary assumption it makes is that the link function is monotone increasing (i.e., that the probability of success on an item strictly increases with ability). This is a very modest assumption that is likely to hold for the sections of CDIs we are concerned with [though it may not hold for morphological over-generalization errors; (Cazden, 1968)]. Second, it was designed to tease apart highly correlated dimensions, such as subsections of ability tests, making it particularly well suited to the testing the grammatical complexity and vocabulary subsections of CDIs which are often correlated at ~ .8 (Bates & Goodman, 1997).

The present study

The present study examined the question of the separability of lexical and grammatical knowledge in a way that accounted for the two challenges outlined above. First, we employed DETECT (Stout et al., 1996) to test the dimensionality of vocabulary and grammar data in a way that is robust to model misspecifications (including linearity), thereby enabling us to determine whether the relationship between the underlying constructs (true vocabulary and true grammar) is more plausibly explained as the product of one system or of two. Second, in addition to testing data from a large sample of children with a broad range of ages, we tested data from two longitudinal samples of the same children assessed at several time points between 16 and 30 months. In doing so, we were able to test whether there were developmental changes in the separability of lexical and grammatical knowledge during this critical time period in language development, as predicted by the mutualism hypothesis. As such, we provide one of the most comprehensive examinations of the separability of lexical and grammatical knowledge to date.

Study 1

Study 1 examined the dimensional structure of vocabulary and grammar using three datasets from the Wordbank American English Words and Sentences data set (downloaded on April, 11th, 2022; Fenson

et al., 2007; Fernald et al., 2013; Marchman et al., 2004; Thal et al., 2013). We first considered all the observations in the American English subsample that were not from longitudinal studies and for which all vocabulary items and all complexity items were available ($N = 2788$, median age =22 months, Age Range = 16: 30 months[2]). We then broke up the remaining data from longitudinal studies into two components: time point 1 ($N = 653$, median age = 17 months, Age Range = 16: 17 months) and time point 2 ($N = 653$, median age = 27 months, Age Range = 27: 28 months), and analyzed these components separately. We analyzed them separately for two reasons: (i), to remove dependent observations from our sample, and (ii) to examine whether the dimensional structure of vocabulary and grammatical knowledge changed during this window.

Method

Materials

We used three subsections of the MBCDI Words and Sentences: 1) the Vocabulary subsection in which parents check off each of the 680 vocabulary items that their child already produces; 2) the combining words question (Has your child begun to combine words yet, such as "nother cracker", or "doggie bite?"), and 3) the grammatical complexity subsection, which contains 37 pairs of items, the more complex of which contains one morphological or syntactic feature that the other less complex item lacks (for example, the second item in the pair *these my tooth* and *these my teeth* contains the correct plural marking and is treated as more complex). Parents check off which of the two items is more characteristic of their child's speech. In practice, parents sometimes choose to select neither option, presumably because the simpler option is still more complex than their child's current syntactic constructions. As validity and reliability evidence reported in the MBCDI manual presupposes coding complex as 1 and simple/non-response as 0 (Fenson et al., 2007) and it is used by Wordbank, we follow the same coding here and in Study 2.

Data preparation

To ensure all data came from children who were already producing two-or-more word sentences, we selected only children whose parents answered 'often' or 'sometimes' to the combining words question (see above). This resulted in subsamples of 2188 children for the primary dataset, 216 children for the 16-month data set (out of 653 total children in the longitudinal data set) and 649 children for the 28-month data set (out of 653 total children).

In order to ensure that we were comparing two sets of items that were unambiguously lexical or grammatical, we included only a subset of the items on the MBCDI. For the Vocabulary section, we included only items that were categorized as nouns, verbs or adjectives (488 of the 680 items). We excluded other items because there are both theoretical and empirical reasons to question whether closed-class items such as prepositions and question words are more appropriately characterized as lexical or grammatical (Braginsky et al., 2019; Day & Elison, 2022). This excluded the subsections Connecting Words, Games and Routines, Helping Verbs (auxiliaries), Locations, People, Places, Pronouns, Quantifiers, and Question Words, Time and Sounds.[3] To measure grammar, we included all items on the Grammatical Complexity subscale. While it would have been possible to also include items from the Word Forms and Word Endings sections, which contain irregular and overgeneralized past tense forms (and noun plural forms), we excluded these items for two reasons. First, it is not clear whether children's early use of morphological knowledge, especially of irregular forms, is best characterized as a lexical or grammatical phenomenon (e.g., see Bybee, 1995; Dabrowska, 2004). Second, it is possible that performance on such items relates to grammatical proficiency in a non-monotone manner, since English-acquiring children exhibit a U-shaped developmental trajectory on irregular past tense forms (Cazden, 1968). This non-monotonicity would violate the assumptions of DETECT, which assumes that higher proficiency in a given construct causes higher performance on a given item (even if the relationship between the construct and item is not linear).

Analytic strategy

In order to test the dimensionality of the vocabulary and grammar items, we used DETECT. We first estimated the one-dimensional composite score using the widely used two-parameter logistic IRT model to a set of vocabulary and complexity items (See Frank et al., 2021 for precedent of using the 2PL model with data from the MBCDI). We then used this composite score to calculate confirmatory DETECT indices, with vocabulary items hypothesized to reflect a single cluster and complexity items hypothesized to a second cluster. While DETECT should, in principle, be insensitive to the choice of smoothing to calculate the composite score, we estimated the composite scores using two additional IRT models as a robustness check. In particular, we also estimated the composite scores using an Empirical Histogram 2PL model, which does not assume that the ability scores are normally distributed, and a spline model which fits non-parametric item response functions. We then calculated the DETECT index again using these composite scores as well. As these two approaches generally yielded similar results, we report on them in Appendix A. As the spline IRT model occasionally produced different results, we discuss these in Appendix B.

DETECT quantifies multidimensionality in a set of items by averaging residual covariances between pairs of items after shared variability has been extracted. If vocabulary and grammar items represent two distinct clusters, then after removing shared variability reflected in the IRT scores described above, items in the same hypothesized cluster (e.g., vocabulary) should positively co-vary with one another. This is because the latent variable, reflecting shared variability across the clusters, should not perfectly measure items within either cluster, resulting in positive residual covariation between items within the same cluster. However, pairs of items from different clusters should be negatively correlated because their shared variability has been removed. DETECT calculates the residual covariances between all pairs of items and averages over (a) 1 times the covariance of each pair of items within a cluster and (b) −1 times the covariance of each pair of items in different clusters. This average reflects how more similar items in the same cluster are than items in different clusters. This measure is an effect size indicating how much larger the conditional covariances are than would be expected in a unidimensional model. Jang and Roussos (2007) offer the following interpretational guidelines for the DETECT index, which they used in their study of Test of English as a Foreign Language (these values are also cited in the *sirt* technical documentation Robitzsch, 2022):

< 0.2: Essential unidimensionality (lexical and subscales likely to reflect the same underlying system/tap the same dimension of linguistic proficiency)[4].

0.2 – 0.4 Weak to moderate multidimensionality.

0.4 – 1.0 Moderate to strong multidimensionality.

> 1.0 Strong multidimensionality.

[2] Not all records in Wordbank contain data from the grammatical complexity section, so this number differs from those reported in other studies.

[3] Note that People and Places contain many words that are proper nouns (e.g., *mummy*) or are part of proper nouns (e.g., *park)* and are coded as "other" rather than "noun" or "predicate" on Wordbank. We follow this coding scheme here.

[4] With unidimensional data, the DETECT index can take negative values when unidimesnionality holds (See, *sirt,* documentation (Robitzsch, 2022).

Table 1
Example DETECT Indices from Previous Studies and Technical Documentation.

Essential Unidimensionality	Simulated 1 dimensional data ($D = -.18$)[a]
$< .2$	Test of English as a Foreign Language (TOEFL): Spoken and Written Expression vs Reading Comprehension ($D = .15$:.17)[b]
	TOEFL Reading Comprehension: Form A vs Form B ($D = .17$)[b]
Mild Multidimensionality	Literacy Assessment and Monitoring Program: Comprehension of Prose vs Documents vs Numerical Texts ($D = .25$: .34)[d]
.2 –.4	Addition and Subtraction Problems ($D = .22$)[c]
	TOEFL: Listening Comprehension vs Reading Comprehension/Spoken and Written Expression (D = .31: .37)[b]
Moderate Multidimensionality	No examples found
.4–1	
Strong Multidimensionality	Big 5 Personality (D = 1.26[a,e])
> 1.0	

[a] See vignettes on conf.detect in *sirt* package (Robitzsch, 2022).
[b] Jang and Roussos (2007).
[c] Analysis conducted on *zareki* dataset from the R package *MPsychoR* (Mair, 2020). Original data from Koller and Alexandrowicz (2010).
[d] Zhang (2013).
[e] Original source (Dolan et al., 2009).

For further context, we have included example DETECT indices from Jang and Roussos (2007), the technical documentation of *sirt,* and other data sources in Table 1. For the sake of exposition, we refer to the above thresholds as *essential unidimensionality*, *mild multidimensionality*, *moderate multidimensionality* and *strong multidimensionality*.

Data from MBCDI differ from educational testing data in two important ways: (i) the relevant subsections of MBCDI are strongly imbalanced and (ii) mean performance on the grammar section of the MBCDI at the earliest age points is very low. We discuss each of these issues in turn. There are 488 open-class items on the MBCDI but only 37 grammatical complexity items. If DETECT were run on the full dataset, the composite score would be greatly influenced by the large number of open-class items. As a result, it would very accurately predict the vocabulary items and their residual covariances would be low. These low covariances would then swamp the DETECT index, which averages across covariances in every pair of items in the dataset. As a result, with large imbalance, the DETECT index is strongly biased toward unidimensionality. While this has been discussed in the methodological literature (see discussions in Jang and Roussos (2007) and Kim (1994)), we are unaware of published solutions.

Therefore, rather than including all vocabulary items in our analysis, we created 100 random subsamples of 37 vocabulary items, sampling from the 488 words without replacement. For each subsample, we (a) fitted three one-dimensional IRT models to the 37 vocabulary and 37 complexity items, and (b) used the resulting proficiency scores to calculate the DETECT index. For a schematic overview, see Fig. 2. Including just a subset of vocabulary items is justifiable for several reasons. First, correlations between scores on vocabulary subsamples and remaining vocabulary items were quite high (typically above .95 and rarely below .9; we report the mean, maximum and minimum value for each analysis). Second, IRT models assume that items are interchangeable and do not affect the meaning of the latent variable. Indeed, many educational tests sample items from large banks, or contain several parallel forms, which are presumed to measure the same construct.

Additionally, mean performance on many items at the youngest time points (Study 1 [16 months] and Study 2 [18/19 and 21 months]) is very low, which creates two problems. First, if no children (or all children) produce a given item, that item does not vary and cannot be used in either IRT (for calculating composite scores) or DETECT. We therefore dropped items that were not produced by any children (or were produced by every child) from the relevant analyses. Second, even if all non-produced items have been removed, low means entail low covariances when data are binary (means and variances are not separable quantities with binary data, so low means imply low variances which imply low covariances). Since DETECT is an average of conditional covariances, extremely low means will deflate the DETECT index. Therefore, when we observed DETECT indices close to or below .2 at the earliest time points, we re-ran the analysis without items that fewer than 5 children produced to see if this unidimensionality was an artifact of relatively low means. We report both sets of results for completeness. Because removing items with particularly low means results in data missing not at random, we encourage caution in interpreting these findings but emphasize that, with sufficiently low means, it becomes mathematically impossible for DETECT to reveal multidimensionality.

Results

Primary data

For all datasets we report the correlation between scores on the vocabulary and complexity subsections for comparison. In the primary Wordbank dataset, this correlation was .84. After removing participants whose parents reported they were not combining words, our sample contained 2188 participants,[5,6]. As can be seen in Table 2, none of the grammar section items were near floor or ceiling for this sample. We then created 100 subsamples of 37 vocabulary items (average correlation between subsample and remaining items = .984, Range = .977: .988), combined these data sets with the complexity items, fit the 2PL model to each, and extracted the resulting latent variable. Using these latent variables, we calculated confirmatory DETECT (hypothesizing that the 37 grammar items and 37 vocabulary items clustered separately). Fig. 3 plots the DETECT index for each data set and each IRT model.

As can be seen in Fig. 3, composites from the 2PL model revealed moderate multidimensionality (average DETECT index = .5, with 1% of DETECT indices $< .4$). In other words, residual covariances between items in the same cluster were larger than residual covariances between items in different clusters, indicating that vocabulary and grammar items appear to tap separable dimensions of linguistic proficiency by the second year.

Longitudinal data (16 months)

After removing participants whose parents reported they were not combining words, the sample contained 216 participants. The correlation between the vocabulary and complexity subscores in this dataset was .411, owing to very low scores on the complexity scale (Median =

[5] There are some participants who contributed more than one data point to the Wordbank data set, even when we select only cross sectional studies. These participants can be potentially identified by a variable in the Wordbank dataset (though the help files note that these may be unreliable). Therefore, we re-ran analyses on a subset of participants who only contributed a single session according to this variable, and observed qualitatively similar results.
[6] After removing participants who were not combining words, the median age was 24 months.

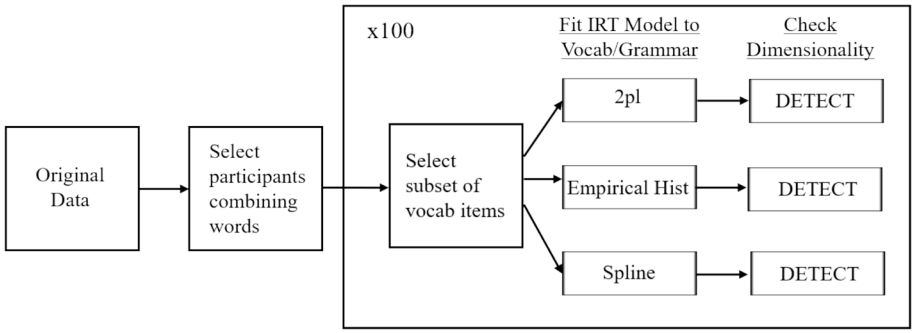

Fig. 2. Data processing and analysis steps for Study 1. For each dataset, we considered all open-class vocabulary items and all grammatical complexity items. We then created 100 datasets of vocabulary items. We merged each of these data sets with the grammatical complexity dataset, estimated the composite scores using one-dimensional item-response theory models and assessed dimensionality using DETECT. Note that results based on composites estimated using the 2PL model are presented in the main manuscript and other results are presented in the appendices.

Table 2
Number of Participants Producing The More Complex Utterance for Each Complexity Item in Study 1.

Item	Primary Dataset	28 Months	16 Months
Two shoe/two shoes	1229	549	22
Two foot/two feet	1173	534	18
Daddy car/daddy's care	1237	546	22
Kitty sleep/kitty sleeping	1167	555	8
I make tower/I making tower	658	381	1
I fall down/I fell down	503	288	5
More cookie/more cookies	1047	510	11
These my tooth/these my teeth	1261	565	17
Baby blanket/baby's blanket	1077	525	12
Doggie kiss me/doggie kissed me	524	335	3
Dady pick me up/daddy picked me up	486	306	0
Kitty go away/kitty went away	357	244	2
Doggie table/doggie on table	976	528	5
That my truck/that's my truck	784	435	10
Baby crying/baby is crying	453	286	3
You fix it/can you fix it	286	182	0
Read me story Mommy/read me a story mommy	456	318	1
No wash dolly/don't wash dolly	589	363	3
Want more juice/want juice in there	572	351	2
There a kitty/there's a kitty	892	476	13
Go bye-be/wanna go bye-bye	744	434	8
Where mommy go/where did mommy go	507	333	3
Coffee hot/that coffee hot	538	366	0
I no do it/I can't do it	691	408	5
I like read stories/I like to read stories	495	321	1
Don't read book/don't want you read that book	507	351	1
Turn on light/turn on that light so I can see	268	216	0
I want that/I want that one you got	316	205	0
Want cookies/want cookies and milk	569	361	1
Cookie mommy/cookie for mommy	671	428	4
Baby want eat/baby want to eat	733	442	5
Lookit me/lookit me dancing	722	437	5
Lookit/lookit what I got	670	393	3
Where's my dolly/ where's my dolly name Sam	304	216	0
We made this/me and Paul made this	446	275	0
I sing song/I sing song for you	404	288	0
Baby crying/baby crying cuz she's sad	446	317	0
N	2188	653	216

0). As can be seen in Table 1, there were nine complexity items for which no child was producing the more complex form and a further 13 for which fewer than 5 children were producing the more complex form. We therefore ran the analysis twice: once considering all items that at least 1 child produced (28 items) and once considering all items that at least 5 children produced (15 items). For each analysis, we created 100 subsets

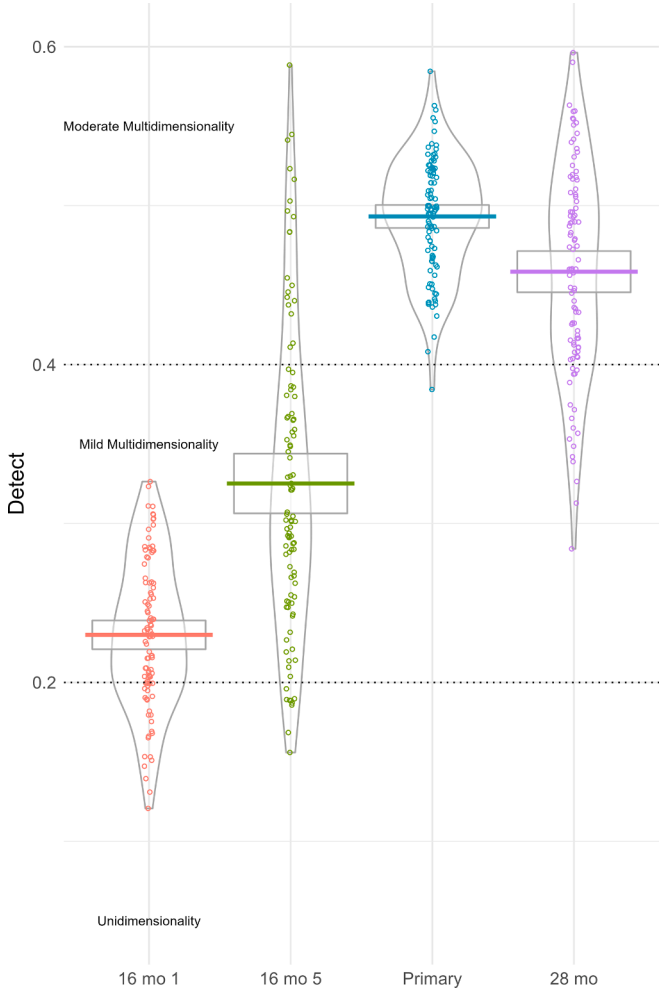

Fig. 3. Pirate plots of DETECT indices each dataset in Study 1. Because of imbalance in the number of items for each hypothesized dataset, we created 100 subsamples of vocabulary and calculated DETECT for each subsample. Each point represents the DETECT index from a given data set, ordered by average age. The X axis indicates the data source. Primary refers to the non-longitudinal data from Wordbank (average age = 24 months). The remaining data sets refer to a large longitudinal dataset on Wordbank, with children tested at 16 and 28 months. At 16 months, there were several items that were produced by fewer than 5 participants. We, therefore, ran our analyses twice, once including (1) and once excluding (5) these items.

of 28 and 100 subsets of 15 vocabulary items, using the same selection criteria as for the complexity items (i.e., selecting vocabulary items that at least 1 or 5 children were producing). When all non-0 observations were included (average correlation between subsample and remaining items = .925, *Min* = .871, *Max* = .950), there was mixed evidence of mild multidimensionality. As can be seen in Fig. 3, the average DETECT index for the 2PL model was .23, with 27% of datasets producing DETECT indices in the essential unidimensionality range. When only items that at least 5 participants produced were considered (average correlation between subsample and remaining items = .89, *Min* = .83, *Max* = .92), there was clearer evidence of mild multidimensionality: As can be seen in Fig. 3, the average DETECT index was .33, with 9% of DETECT indices below .2.

Longitudinal data (28 months)

After removing participants whose parents reported they were not combining words, this sample contained 649 participants ($r_{vocab,grammar}$ = .777). As can be seen from Table 1, there was no evidence of ceiling or floor effects for any grammar items. As a result, we ran the same procedure as for the cross-sectional data above (average correlation between subsample and remaining items = .969, Range = .950: .978). As can be seen in Fig. 3, the 2PL model suggested moderate multidimensionality with an average DETECT index of .46 (and 17% of DETECT indices < .4).

Interim summary

Contrary to lexicalist approaches, we observed clear separation of vocabulary and grammar in our primary data set, suggesting that vocabulary and grammatical knowledge are partially separable by the second year. We observed mixed evidence for the separability of lexical and grammatical knowledge at 16 months, but this was probably due to the very low scores on the complexity items in this age group. When only items that at least 5 participants produced were analyzed, we saw clearer evidence of two distinct dimensions. Consistent with the results from the primary dataset, we observed clear evidence of the separability of lexical and grammatical knowledge at 28 months.

While Study 1 provided evidence in favour of multidimensionality across the primary dataset and longitudinal samples, the longitudinal data in Wordbank contained only two time points, limiting the possibility of observing different patterns across time points. In Study 2 we analysed a more intensively sampled longitudinal data from two additional dialects of English – British English and Australian English. Thus, with these new data we tested whether the pattern of results observed in Study 1 replicated in two more intensely sampled cohorts, thereby also testing the generalisability of the results across different dialects of English, which do not always align in some features of acquisition (e.g., Cattani et al., 2019).

Study 2

In Study 2, we used data from two large-scale longitudinal studies of language acquisition and processing, the Language 0–5 Project (L05; Rowland et al., 2018; Peter et al, 2019) and the Canberra Longitudinal Child Language Project (CLCL; Donnelly & Kidd, 2021; Kidd et al., 2018). In both studies, adaptations of the MBCDI Words and Sentences were administered at several time points (18, 21, 24 and 30 months in CLCL and 19, 21, 24, 25 and 30 months in L05). This allowed us to track the dimensionality of vocabulary and grammar in the same children at four different time points (18/19 months, 21 months, 24 months and 30 months).

Method

Participants

CLCL

Participants in CLCL participated in a set of laboratory-based tasks and home recordings several times a year from 9 months to 5 years (every three months from nine to twenty-four months and every six months from twenty-four to sixty months). Inclusion criteria for the study were: (i) full-term (at least 37 weeks gestation) babies born with a typical birth weight (> 2.5 kg), (ii) a predominantly monolingual language environment (mean percentage of language other than English = 2%, Range: [0, 40%], Mode = 0), and (iii) no history of medical conditions that would affect typical language development, such as repeated ear infections, visual or hearing impairment, or diagnosed developmental disabilities. Participants were recruited from Canberra, a medium-sized Australian city with high socio-economic status. Approximately 75% of the parents had completed a bachelor degree or higher and ethnicity information was not recorded.

Of the 130 participants who were initially recruited for the study, nine had dropped out by the 18-month session. As a result, 121 participants are included in the present analysis. Of these 118, 115 completed the 18-month session, 113 completed the 21-month session, 116 completed the 24-month session and 100 completed the 30-month session. Of these, 93 completed all 4 sessions, 21 completed 3 sessions, 3 completed 2 sessions, 3 completed 1 session.

L05

Participants in L05 participated every three months from 9 to 24 months and then every six months from 24 to 54 months of ages. Families were recruited from the north of England, UK. Inclusion criteria for the study were (i) full term babies with typical birth weights and (ii) no evidence of atypical development. Of the 95 participants originally recruited, 90 were still participating at 19 months. Of the 90 participants, 88 were White British, 59 reported a university degree or higher and 62 reported an annual (before tax) income of £42,001 per year or more (i.e. above the higher rate tax bracket threshold at the time of recruitment).

Of the 90 participants, 83 completed at least 1 vocabulary checklist, 79 at the 19-month session, 82 at the 21-month session, 75 at the 24-month session and 74 at the 30-month session. In total, 68 completed all four sessions, 9 completed three sessions, 5 completed 2 sessions and 1 completed 1 session.

Instruments

CLCL

Parents completed the (US English) MBCDI—Words and Sentences form when their children were between 18 and 30 months. To better capture the Australian dialect, some minor changes were made to a small number of words which resulted in a total of 678 items (Reilly et al., 2007).

L05

Children completed the (UK English) Lincoln CDI (Meints et al., 2017), an adaptation of the MBCDI to UK English, which contains 689 words.

Combining the instruments

Because these two instruments are adaptations of the MBCDI to the local linguistic context, their open-class vocabulary sections do not entirely overlap. Items that differed in spelling or were near synonyms across the two forms were matched and treated as the same lexical item (see Appendix C for list for a list of items matched across the two data sets). Items with no near synonym on the two lists were dropped (26 from CLCL and 36 from L05), resulting in 653 in total.

Table 3
Number of Participants Producing More Complex Utterance for Each Complexity Item in Study 2.

Item	18/19	21	24	30
Two shoe/two shoes	26	64	111	146
Two foot/two feet	18	57	93	134
Daddy car/daddy's care	18	71	123	152
Kitty sleep/kitty sleeping	14	61	110	146
I make tower/I making tower	6	24	74	125
I fall down/I fell down	2	7	38	115
More cookie/more cookies	1	3	22	93
These my tooth/these my teeth	2	7	41	112
Baby blanket/baby's blanket	1	14	36	120
Doggie kiss me/doggie kissed me	5	14	55	105
Daddy pick me up/daddy picked me up	9	29	75	135
Kitty go away/kitty went away	2	12	55	130
Doggie table/doggie on table	2	8	46	118
That my truck/that's my truck	4	18	45	100
Baby crying/baby is crying	11	41	81	142
You fix it/can you fix it	26	65	108	155
Read me story Mommy/read me a story mommy	12	51	107	145
No wash dolly/don't wash dolly	4	15	40	113
Want more juice/want juice in there	2	6	34	115
There a kitty/there's a kitty	1	9	52	131
Go bye-be/wanna go bye-bye	1	3	28	104
Where mommy go/where did mommy go	0	6	41	121
Coffee hot/that coffee hot	0	1	10	84
I no do it/I can't do it	0	7	17	100
I like read stories/I like to read stories	2	13	50	128
Don't read book/don't want you read that book	1	21	74	144
Turn on light/turn on that light so I can see	0	8	33	110
I want that/I want that one you got	0	7	22	88
Want cookies/want cookies and milk	5	25	88	148
Cookie mommy/cookie for mommy	7	25	70	133
Baby want eat/baby want to eat	0	11	56	133
Lookit me/lookit me dancing	5	30	73	137
Lookit/lookit what I got	2	17	59	140
Where's my dolly/ where's my dolly name Sam	0	4	22	83
We made this/me and Paul made this	0	8	31	93
I sing song/I sing song for you	0	2	26	111
Baby crying/baby crying cuz she's sad	0	11	45	119
N	104	148	176	170

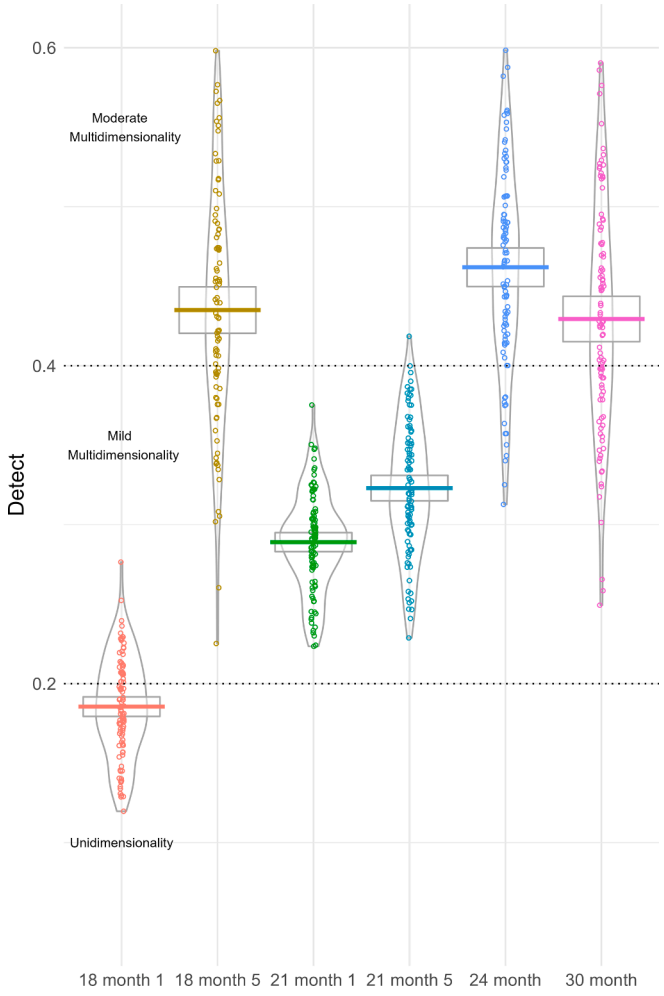

Fig. 4. Pirate plots of DETECT indices from each dataset in Study 2. Because of imbalance in the number of items for each hypothesized dataset, we created 100 subsamples of vocabulary and calculated DETECT for each subsample. Each point represents the DETECT index from a given data set. The X axis indicates the data source. At 18 and 21 months, there were several items that were produced by fewer than 5 participants. We, therefore, ran our analyses twice, once including (1) and once excluding (5) these items.

Results

We followed the same analytic strategy from Study 1 for all four time points in Study 2.

18/19 months

Of the 194 participants, 104 were reported to be combining words and were included in the present analyses ($r_{vocab/grammar} = .65$). As can be seen in Table 3, there were 10 items that no child was producing and a further 14 items that fewer than five children were producing. We thus followed the same strategy as in the younger subsample of Study 1, running our analyses twice, once including all items that at least one child was producing and once including all items that at least five children were producing. As can be seen in Fig. 4, when the 27 complexity items which at least one participant produced were included (mean correlation between included and excluded vocabulary items = .95, Range = .918: .971), the 2PL model revealed evidence of essential unidimensionality (mean DETECT = .186, 34% of DETECT indices > .2). When we only considered the 13 complexity items that at least 5 participants were producing (mean correlation between included and excluded vocabulary items = .92, *Range* = .87: .95), the 2PL model revealed moderate multidimensionality (mean DETECT = .456, 28% of DETECT indices < .4). In sum, while results for all items produced by at least 1 child suggested undimensionality, this may have reflected the relatively low means on some items (and therefore low variances and covariances). When only items that at least 5 participants produced were included, evidence of multidimensionality was very similar to that observed in Study 1.

21 months

Of the 194 participants, 148 were reported to be combining words and were included in the present analyses ($r_{vocab/grammar} = .777$). As can be seen in Table 3, there were 5 items on the complexity scale that fewer than 5 participants were producing. Therefore, we again ran our analyses on both the full grammatical complexity scale and on the subset of items that at least 5 participants were producing. As can be seen in Fig. 4, when all items from the complexity scale were analyzed (mean correlation between vocabulary subsample and remaining items = .969, Range = .951: .978), the 2PL model revealed mild multidimensionality (mean DETECT = .289, 0% of DETECT indices < .2). When only the 32 items which at least 5 participants produced were analyzed (mean correlation between vocabulary subsample and remaining items = .965, Range = .941: .977), the 2PL model revealed mild multidimensionality (mean DETECT = .323 with 0% DETECT indices < .2).

24 months

Of the 190 participants, 176 were reported to be combining words

and were included in the present analyses ($r_{vocab/grammar} = .755$). As can be seen in Table 3, there was no evidence of ceiling or floor effects for any grammar items and the analysis was run on all items (Average correlation between subsample and remaining items = .979, Range = .955: .969). As can be seen in Fig. 4, the 2PL model suggested moderate multidimensionality with an average DETECT index of .467 (and 13% of DETECT indices < .4).

30 months

Of the 174 original participants, 170 were reported to be combining words ($r_{vocab/grammar} = .758$). As can be seen in Table 3, there was no evidence of ceiling or floor effects for any grammar items. As a result, we ran the same procedure as with the cross-sectional data above (Average correlation between subsample and remaining items = .969, Range = .948: .981). As can be seen in Fig. 4, the 2PL model suggested moderate multidimensionality with an average DETECT index of .423 (and 37% of the vocabulary subsets suggesting mild multidimensionality).

Summary

Across the four time points we observed a consistent pattern of multidimensionality. At the earliest age, 18/19 months, we observed a unidimensional structure when all items produced by at least one child on the Grammatical Complexity scale were included. However, this was likely an artifact of the floor effect due to children's limited linguistic proficiency: when we included only items which five or more children produced, we observed clearer evidence of moderate multidimensionality. At 21 months, we observed evidence of mild multidimensionality, a pattern that did not qualitatively change after excluding items that fewer than five children were producing. And at 24 and 30 months, we observed mostly moderate multidimensionality. In sum, Study 2 largely confirmed the findings from Study 1. While grammar and vocabulary are highly correlated, they were separable from a very early age, likely before children's second birthday.

Discussion

In the current paper we addressed a long-standing issue in the language acquisition literature – the degree to which early vocabulary and grammar are separable, if at all. The issue has proven particularly difficult to tackle, owing to a range of analytic problems that have not been adequately addressed. Mindful of these past problems, we used DETECT, a form of non-parametric IRT, to model the dimensionality of early vocabulary and grammar using several large samples of children whose language was measured via parental report. Using data from Wordbank (Frank et al., 2017), we found clear evidence that, contra lexicalist approaches (Bates & Goodman, 1997), vocabulary and grammar were clearly separable before children's second birthdays. Moreover, using longitudinal data from both Wordbank and two more extensively sampled longitudinal studies of language acquisition, we found clear evidence that the two constructs exhibited moderate multidimensionality across most time points. Taken together, these results suggest that, while the correlation between vocabulary and grammar is impressively strong, the two are separable, at least within the range of grammatical proficiency measured by the CDIs.

These results show that items on the vocabulary and grammatical complexity sections of the MBCDI and Lincoln CDI are statistically separable in a confirmatory manner. On their own, latent variable models can only tell us that items reflect separate dimensions but cannot tell us why this is. If the two sections of the instruments induce different response biases, with some parents overestimating their child's vocabulary and some parents overestimating their child's grammar, this could create multidimensionality. However, we argue the multidimensional structure reflects differences in the content of these subtests (vocabulary vs grammar) rather than differences in responses biases for several reasons. First, as the MBCDI manual (Fenson et al., 2007) notes, the subsections of the MBCDI Words and Sentences have excellent validity. Notably, the shape and variance of the developmental trajectories of these subsections closely match those observed in observational studies. For example, the rank order of acquisition of grammatical items closely matches what is observed in observational research (e.g., noun inflections are acquired before verb inflections), and the scores of each subsection correlate strongly with concurrent measurements in laboratory and observational studies. Second, the correlation between scores on the grammatical complexity and productive vocabulary subscores (on the MBCDI: WS) actually exceeds the correlation between productive and receptive vocabulary (on the MBCDI: WG), which is difficult to explain if some parents are more attuned to their child's vocabulary or grammar (Bates & Goodman, 1997). Third, the correlation between the word forms section and vocabulary subsections of the MBCDI ($r = .82$ on Wordbank) is nearly identical to the correlation between the grammatical complexity and vocabulary subsections ($r = .84$ in our largest subsample), even though the response format of the word forms section is identical to that of the vocabulary section (it asks parents to indicate whether their child produces an inflected form rather than asking to choose which example is more like their child's speech). Taken together, these considerations provide strong evidence that the source of multidimensionality is the content in the two sections and not a methodological artifact.

What do these results reveal about the differences between vocabulary and grammatical knowledge? We emphasize at the outset that separable does not mean completely distinct. Rather, our data show that despite significant overlap between vocabulary and grammatical knowledge (with correlations as high as .84), the two have qualities that make them irreducible to one dimension and exhibit mild to moderate multidimensionality. We take this as consistent with both linguistic theories and psycholinguistic data. Nearly all formal theories of grammar assume important interactions between lexical and grammatical knowledge, whether assuming lexical entries for words contain rich grammatical information (Bresnan 2001; Chomsky, 1995; Pollard & Sag, 1994; Steedman, 2000), or that both words and syntactic rules are stored items that differ in their degree of productivity (Jackendoff, 2013) or abstractness (Goldberg, 1995). Moreover, psycholinguistic research shows that speakers store multi-word strings (Arnon et al., 2017; Bannard & Matthews, 2008; Arnon & Snider, 2010; Jolsvai et al., 2020; Snider & Arnon, 2012) and that these play a key role in acquisition and processing (Abu-Zhaya et al. 2022; McCauley, et al., 2021). Therefore, we should not be surprised that vocabulary and grammar are so highly correlated.

Although vocabulary and grammar are intimately related, our results show that they can be partially separated, in a confirmatory manner, from a quite young age. In many of our analyses, we observed DETECT indices that exceed those reported by Jang and Roussos' (2007) study of the listening and reading comprehension/SWE sections of the ToEFL, suggesting closely related yet separable domains of knowledge (See Table 1). Even in our youngest samples, we observed evidence of mild to moderate multidimensionality (at least when removing items with counts less than 5), despite the low means biasing these indices toward unidirectionality. These results suggest that that the grammatical complexity and vocabulary sections of English CDIs measure different, though certainly overlapping, domains of knowledge. Unlike lexical items on the vocabulary subsection, the items on the grammatical complexity section of English CDIs mostly concern properties such as agreement, morphology, auxiliary verbs, prepositions and subordinate clauses. In other words, the complexity section measures children's use of formal operators and processes that can be applied to large classes of words. Our results demonstrate that English-speaking two-year olds' knowledge of the formal aspects of their language is not entirely determined by their vocabulary knowledge.

One important caveat is that our approach did not allow us to study the earliest stages of grammatical development. A number of children

were reported to be combining words, but then scored zero on the complexity section of the MBCDI ($N = 464$ out of 2188 participants in the Wordbank sample). This suggests that there are earlier stages of grammatical development that are not captured in our analyses. This is not surprising, given that the complexity section of the CDIs was designed to measure early formal features that follow children's earliest two- to-three-word productions. It is possible that children's earliest two-to-three-word productions are completely lexical, consistent with some usage-based accounts of grammatical development (Bannard et al., 2009; Dabrowska & Lieven, 2005; Lieven, Pine & Baldwin, 1997).

However, our results reveal earlier grammatical knowledge than is typically reported in these studies. For example, based on the results of a computational model of children's production data, Bannard et al. (2009) argued that children do not rely on abstract grammatical representations until sometime between their second and third birthdays. This may reflect the limitations of corpus data traditionally used in these studies. Because children rarely use advanced grammatical constructions in spontaneous speech, child language corpora are often too small to reliably detect children's grammatical competence (Meylan et al., 2017). Indeed, experimental studies often show evidence of grammatical competence at earlier ages than corpus studies (Ambridge et al., 2008; Gertner et al., 2006; Messenger & Fisher, 2018; for *meta*-analytic evidence see Cao & Lewis, 2022). Parental checklists, like the CDI, may provide better estimates of the upper limits of children's grammatical abilities than do corpus studies. This may be in part due to sampling problems with corpora (Rowland & Fletcher, 2006). Meylan et al (2017) developed a Bayesian computational cognitive model to test whether children's earliest use of determiners was better explained by item-specific or grammatical knowledge. While they found that most corpora did not contain enough data to distinguish between the two possibilities, they observed that data from one extremely dense corpus, the Speechome corpus (Roy et al., 2015), showed an initial phase of item-specificity followed by rapid increase in the degree of abstractness shortly after the onset of grammatical productivity. This encourages caution when extrapolating our results to levels of grammatical knowledge below what can be measured by CDIs and motivates the development of measures that may be more sensitive to emergent grammatical knowledge in this window.

Whatever the nature of children's grammatical knowledge at these earliest levels, our results provide clear evidence of the separability of lexical and grammatical knowledge in English-speaking children by their second birthdays. This sets important boundaries on theoretical development in first language acquisition, where there is significant debate regarding the origin and nature of grammatical knowledge. The debate has raged in various guises across different domains (e.g., inflectional morphology, Marcus, 1992; Marchman & Bates 1994; grammar, Tomasello, 2000; Fisher, 2002). Lexicalist approaches have emphasized common representational principles and learning mechanisms, whereas dual process approaches have typically emphasized the separability of the lexicon and grammar, typically aligning themselves with formal approaches to grammar. While the specific predictions of many accounts depend on a number of assumptions about the interactions between abstract linguistic representations and the lexicon, our data rule out a single mechanism lexicalist explanation for early grammatical development, at least by the second year, just as the strong correlation between the two systems (Bates & Goodman, 1997) and clear evidence for the role of stored multi-word utterances in acquisition and processing rule out the strictest dual-systems accounts (Abu-Zhaya et al. 2022; Arnon, et al., 2017; Arnon & Snider, 2010; Bannard & Matthews, 2008; Jolsvai, et al., 2020; McCauley, et al., 2021; Snider & Arnon, 2012).

When contextualized within the broader field, the finding of separable vocabulary and grammatical systems can sharpen accounts of acquisition. Notably, the result points towards a necessity to postulate an early operable mechanism that isolates and abstracts over multiple exemplars of categories, whether they be grammatical categories (e.g., something akin to NOUN, VERB) or bound morphemes (e.g., *−ed, −s*). In reality, *all* theories of acquisition assume a distributional learning mechanism like this, with even young infants demonstrating abilities to identify regularities over purely formal stimuli via processes like statistical learning (Saffran et al., 1996; see also Aslin & Newport, 2012). Indeed, even proponents of some of the most radically lexicalist computational models in language acquisition argue that such models are improved with the addition of a mechanism that allows some form of generalization (Ambridge, 2020b; McCauley & Christiansen, 2019b). For example, McCauley & Christiansen (2019b) found that the model reported in McCauley and Christiansen (2019a), which models acquisition as the verbatim storage of chunks based on transitional probabilities, was improved when allowed to abstract over individual lexical items and develop slot and frame constructions (e.g., "Here's an XX"). While lexical items are the basic ingredients of this process, the mechanism's output is knowledge that is more than the sum of its parts – a partially lexically-independent grammar (Naigles, 2002).

The outstanding question concerns the cognitive architecture of these systems and their overlap. Many theoretical proposals point to the separability of form and meaning. For instance, Chang et al.'s (2006) Dual path computational model separates syntactic sequencing and meaning in two distinct dynamic pathways, which allows it to generalise well beyond its input (see also Dell & Chang, 2014). Neurologically inspired models like the Declarative/Procedural model divide the labor of syntactic sequencing and lexical representation into different and competing memory systems (Ullman, 2004), whereas others like the Memory Unification and Control model (Hagoort, 2013) postulate dynamic processes that operate over domain knowledge (e.g., 'syntactic' and 'semantic' unification). Indeed, lexical and grammatical processes do partially dissociate behaviorally. For instance, in acquisition abstract and lexically-based syntactic priming have different developmental trajectories (Kumarage et al., 2022; Rowland et al., 2012), and children with Developmental Language Disorder struggle most with morphological processes (e.g., tense marking in English) and less so with lexical learning, which may be due to abnormalities located in the basal ganglia, part of the procedural learning system (Ullman et al., 2024). Our point is not that this provides evidence for rule-based formal syntactic knowledge (*pace* Pinker, 1994), but that the acquisition of formal knowledge must at least partially depend upon on meaning-independent sequencing systems, quickly leading to the development of abstract structural knowledge. Determining the nature of this knowledge requires crucial theoretical development that is sorely needed in the field. Our major contribution is identifying a clear developmental benchmark that any plausible theoretical account must explain.

As a final point, we did not observe evidence of a unidimensional system at the oldest ages (28 months in Study 1 and 30 months in Study 2). While our study was designed to rule out the possibility of a unidimensional solution caused by mutualistic coupling between vocabulary and grammar, our results should not be taken as evidence that there is no mutualistic coupling. Indeed, mutualistic coupling can result in a multidimensional factor structure. For example, Van der Maas et al. (2006) found that their mutualistic coupling model developed a hierarchical factor structure with one primary factor and three subfactors, depending on the pattern of causal relationships between the initially uncorrelated variables. Moreover, it is possible that vocabulary and grammar become more strongly correlated before the school years (LARR, 2015; Monaghan et al., 2023; Tomblin & Zhang, 2007). These possibilities highlight the value of developmental studies in understanding the nature of psycholinguistic and cognitive constructs (Van der Maas et al., 2006; Kievit et al., 2017; Kievit et al., 2019).

Limitations

While our approach reflects an novel solution to a persistent problem in developmental psycholinguistics, it is not without limitations. First, as mentioned above, many children who were reported to be combining

words scored zero on the grammatical complexity section of the CDI, meaning we were unable to study the dimensional structure of vocabulary and grammar in the earliest stages of grammatical development. Additionally, due to the low means on grammatical complexity at the earliest time points (in particular, 16 months in Study 1 and 18/19 and 21 months in Study 2), we removed items that fewer than five children produced. Importantly, results with and without these items sometimes differed (in particular, the 16-month data set from Study 1 and the 18/19 month dataset from Study 2). While these differences are not surprising given the mathematics of DETECT, removing data in a non-random way could introduce bias and we encourage caution in interpreting evidence from the youngest ages. Despite this limitation, results with these low-count items removed were broadly similar (suggesting the mild to moderate multidimensionality) to results from later ages, where low grammatical complexity scores was not an issue. Second, because DETECT was developed in the educational testing literature, it has not yet been applied to data of this sort, and the magnitude of the DETECT index can be difficult to interpret. We have collected indices from published substantive and methodological studies and interpreted our results relative to Jang and Roussos (2007)'s study of the dimensionality of the ToEFL. In the future, domain-specific benchmarks would be more informative. We note, however, that this is a challenge with effect sizes in general, which are difficult to interpret without such domain-specific benchmarks. Third, while our analytic approach was well suited for examining whether vocabulary and grammatical knowledge are multidimensional at different ages, it would not be well suited for explanatory modeling examining how vocabulary, grammar or their correlation vary as a function of other variables. This reflects both a strength and limitation of non-parametric methods like DETECT: by making very general assumptions about the data-generating process, DETECT is compatible with a wide range of data types but is limited in the hypotheses it can test. Future research aimed at developing theoretically motivated parametric latent variable models will provide an important complement to the present approach (Huang & Bolt, 2024).

Conclusion

In sum, our results demonstrate that vocabulary and grammatical knowledge, while highly correlated, are statistically separable from very early in language development, at least by the level of grammatical development measurable with CDIs. This suggests that the acquisition of lexical knowledge and grammatical knowledge are supported by at least partially separable lower-level cognitive processes. Moreover, our results demonstrate the power of developmental designs and psychometrics-inspired models. Future research should aim to understand the separability of vocabulary and grammatical knowledge at the earlier levels of grammatical development and examine the relationship between vocabulary and grammar in languages with richer morphologies, where the distinctions between lexical and grammatical processes is less clear than in English (Kidd & Garcia, 2022).

CRediT authorship contribution statement

Seamus Donnelly: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Evan Kidd:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Jay Verkuilen:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Caroline Rowland:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially supported by The Australian Research Council (CE140100041: CI Evan Kidd), the Economic and Social Research Council (ESRC) International Centre for Language and Communicative Development (LuCiD), funded by the U.K. Economic and Social Research Council (ES/L008955/1 and ES/S007113/1), and the Max Planck Society. We thank Jinming Zhang for statistical advice.

Appendix A:. Analyses using different IRT models

In the analyses in our main paper, we used the 2PL IRT model to calculate composite scores. However, other parametric IRT models are possible for this task. While DETECT should not be greatly influenced by the choice of model used to calculate the composite score, we considered two additional IRT models as a robustness check. First, we considered the 2PL Empirical Histogram model (EH), which like the 2PL model assumes a logistic item-response function but does not require the distribution of participant ability scores to be Gaussian. Second, we considered a spline IRT model, which does not assume a particular functional form for the item-response function and can therefore take a variety of shapes (though it does assume that the participant ability scores are normally distributed).

Wordbank primary data set

Consistent with results from the 2PL model, results from the EH model suggested moderate multidimensionality (See Fig. A1. Average DETECT index = .5, with 1% of indices < .4) as did results from the spline model (See Fig. A2. Average DETECT index = .45, 15% of detect indices below .4 respectively).

Wordbank Longitudinal Data Set

16 months. Consistent with results based on composites using the 2PL model, composites from the Empirical Histogram model suggested weak multidimensionality (Average DETECT index = .21, with 49% of indices below .2). Composites from the spline model produced almost entirely evidence of essential unidimensionality (99% of DETECT indices below .2); however, the spline model fit the data very poorly and predicted a non-monotonic relationship between true and observed scores for many of the datasets, likely because of the relatively small sample size and the flexibility of the spline-based item response function (See Appendix B for more details). When the analysis considered only items for which at least 5 children were producing the more complex from, composites from the EH model suggested weak multidimensionality (average DETECT index = .29, 24% below .2). As was the case when all non-0 items were included, composites from the spline model produced evidence of unidimensionality (average DETECT index = .041, 100% below .2); however, we believe this is due to the misfit of the model (See Appendix B).

28 months. Similar to those from the 2PL model, composite scores based on the empirical histogram and spline models produced evidence of moderate multidimensionality (mean DETECT scores of .464 and .420 respectively, with 17% and 42% of DETECT indices below.4 respectively).

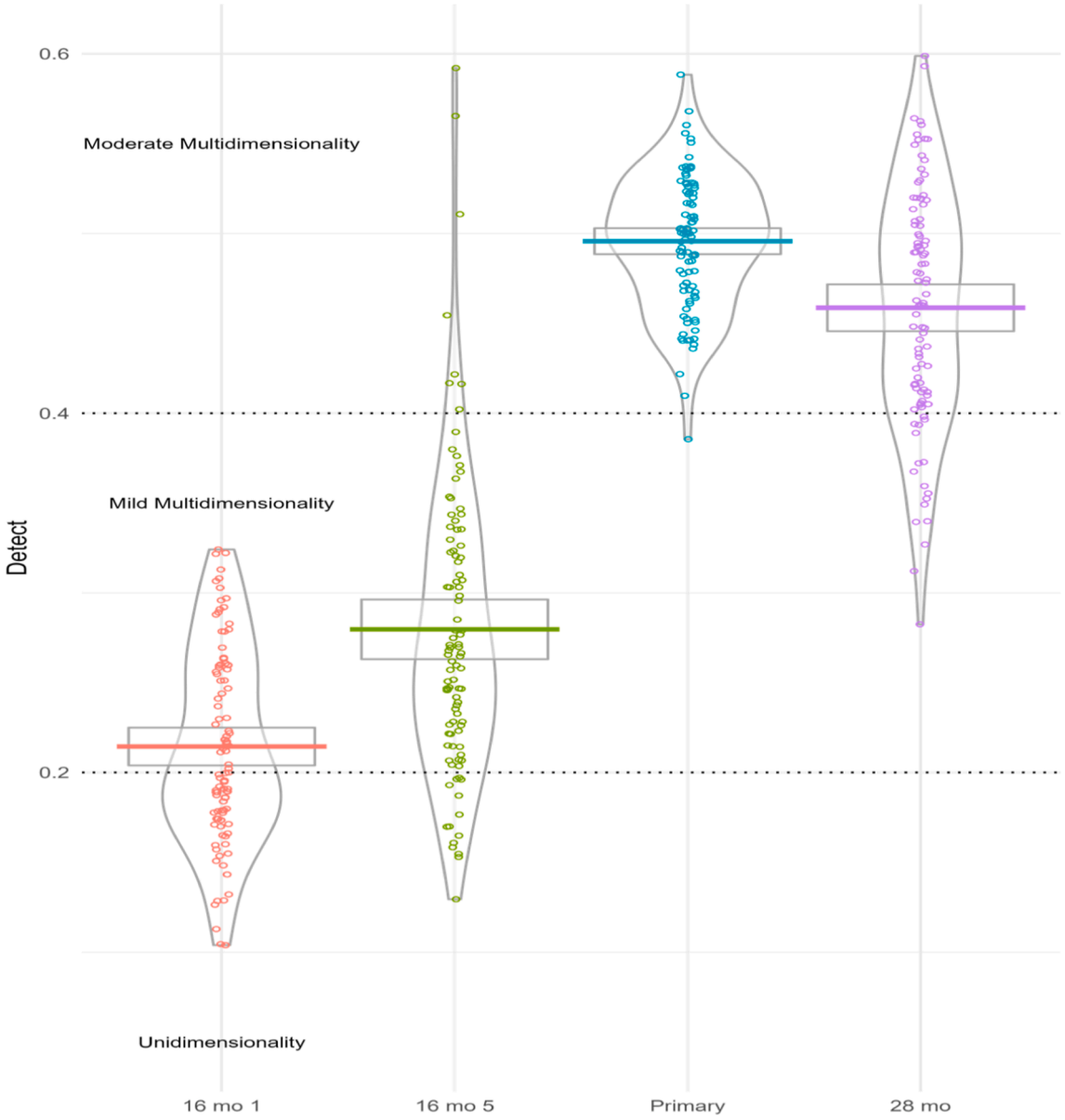

Fig. A1. Pirate plots of DETECT indices from the empirical histogram IRT models in Study 1. Because of imbalance in the number of items for each hypothesized dataset, we created 100 subsamples of vocabulary and calculated DETECT for each subsample. Each point represents the DETECT index from a given data set. The X axis indicates the data source. Primary refers to the non-longitudinal data from Wordbank. The remaining data sets refer to a large longitudinal dataset on Wordbank, with children tested at 16 and 28 months. At 16 months, there were several items that were produced by fewer than 5 participants. We, therefore, ran our analyses twice, once including (1) and once excluding (5) these items.

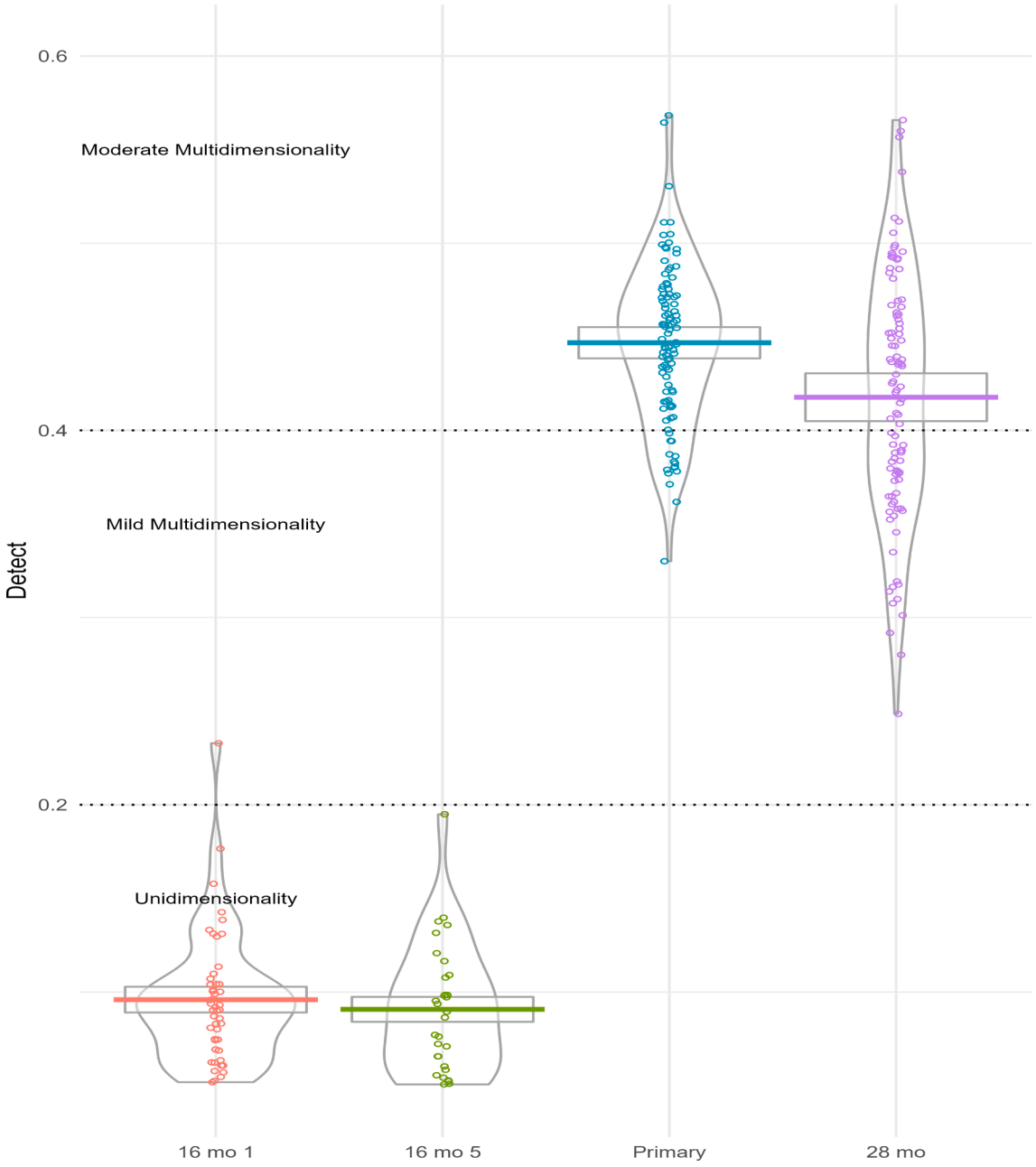

Fig. A2. Pirate plots of DETECT indices from the spline IRT models in Study 1. Because of imbalance in the number of items for each hypothesized dataset, we created 100 subsamples of vocabulary and calculated DETECT for each subsample. Each point represents the DETECT index from a given data set. The X axis indicates the data source. Primary refers to the non-longitudinal data from Wordbank. The remaining data sets refer to a large longitudinal dataset on Wordbank, with children tested at 16 and 28 months. At 16 months, there were several items that were produced by fewer than 5 participants. We, therefore, ran our analyses twice, once including (1) and once excluding (5) these items.

CLCL/L05 Data

18/19 months. When all items that least 1 child was producing were included, composite scores estimated in the empirical historgram and spline models yielded similar results to those from the 2PL model (Fig. A3; mean DETECT = .196 and .139 respectively, 43% and 6%, respectively, of DETECT indices > .2). When only items that at least 5 children were producing were included, composites estimates in the empirical histogram model suggested moderate multidimensionality (mean DETECT = .465, 25% of DETECT indices < .4) whereas those estimate in the spline model suggested weaker multidimensionality (mean DETECT = .334, 77% of DETECT indices < .4).

21 months. When all items that least 1 child was producing were included, composite scores estimated in the empirical histogram and spline models yielded similar results to those from the 2PL model (Fig. A4; mean DETECT = .290 and .295 respectively, with 0% and 1%, respectively, of DETECT indices < .2). When all items that at least 5 children were producing was included, composite scores estimated using the empirical histogram and spline models again yielded similar results to those estimated by the 2PL model (mean DETECT = .328 and .332, respectively with 0% of DETECT indices < .2).

24 months. Composite scores estimated in the empirical historgram and spline models yielded similar results to those from the 2PL model (average

DETECT indices = 465 and.400 respectively with 14% and 51%, respectively, of DETECT indices < .4).

30 months. Composite scores estimated using the empirical histogram model yielded similar results to those from the 2PL model (mean DETECT scores of .44 with 34% of DETECT indices < .4) whereas those estimated with spline model revealed weaker multidimensionality (mean DETECT score of .35, with 18% of DETECT indices > .4).

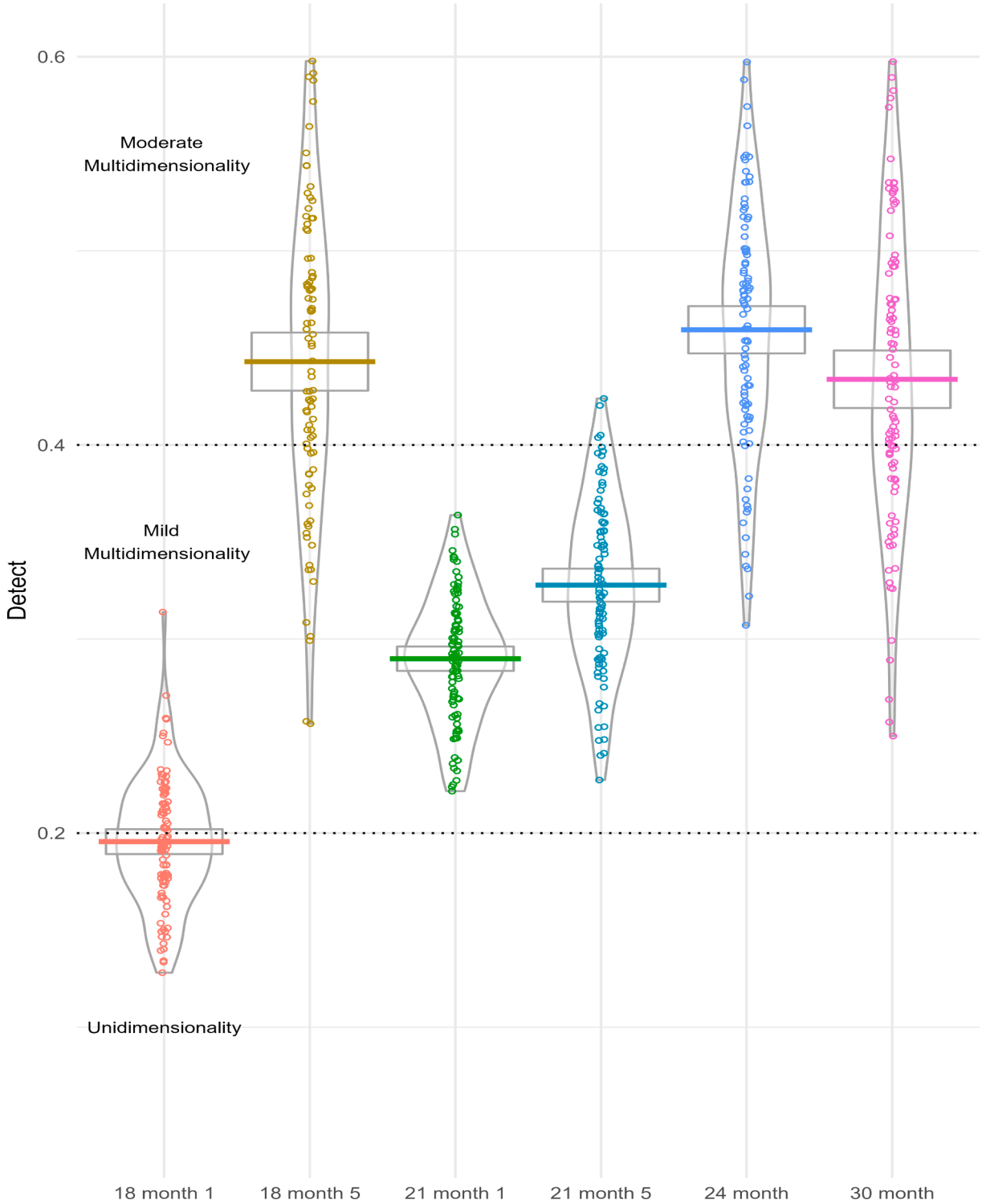

Fig. A3. Pirate plots of DETECT indices from composite scores estimated using empirical histogram IRT models on data from Study 2. Because of imbalance in the number of items for each hypothesized dataset, we created 100 subsamples of vocabulary and calculated DETECT for each subsample. Each point represents the DETECT index from a given data set. At 18 and 21 months, there were several items that were produced by fewer than 5 participants. We, therefore, ran our analyses twice, once including (1) and once excluding (5) these items.

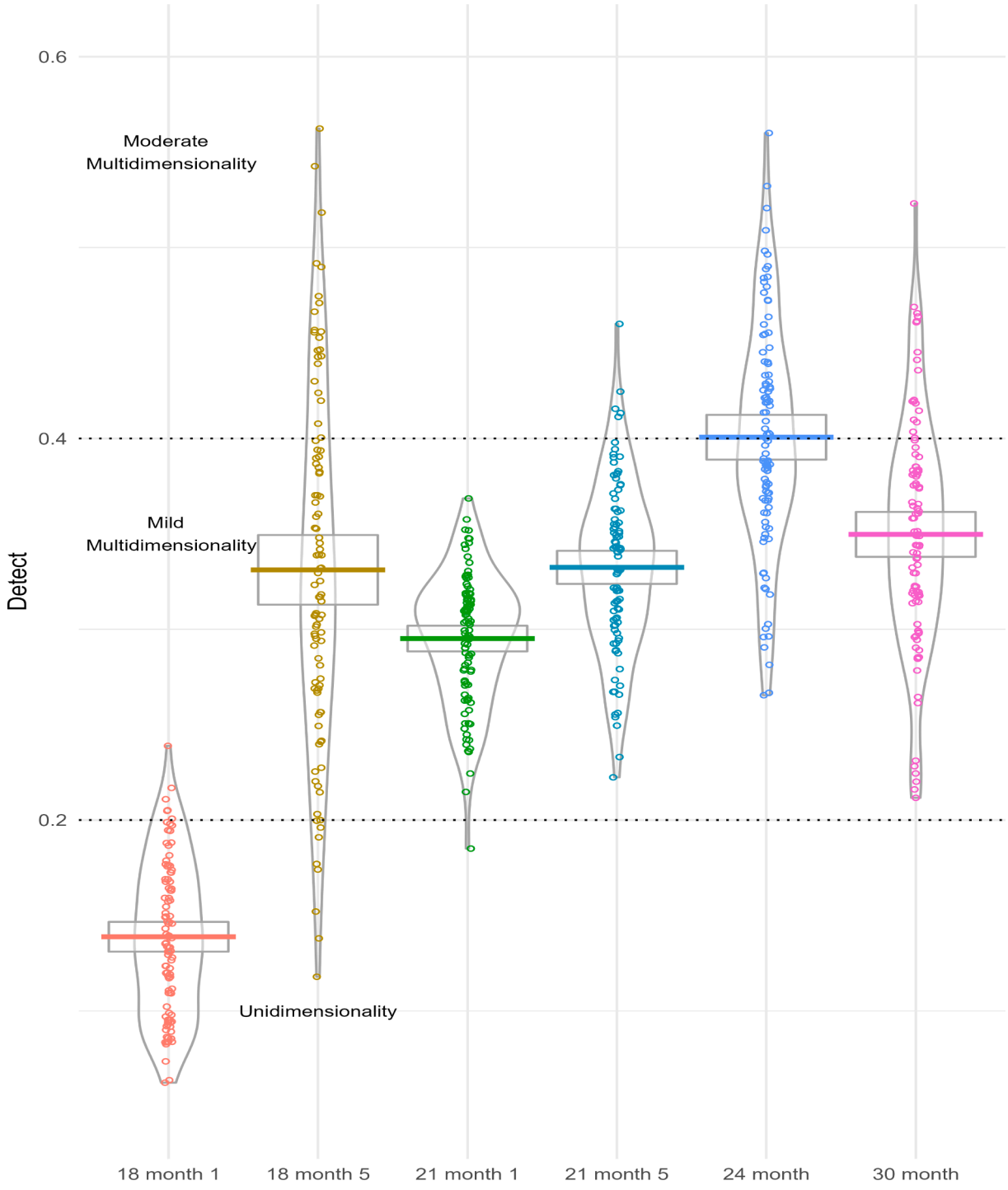

Fig. A4. Pirate plots of DETECT indices from composite scores estimated using spline IRT models on data from Study 2. Because of imbalance in the number of items for each hypothesized dataset, we created 100 subsamples of vocabulary and calculated DETECT for each subsample. Each point represents the DETECT index from a given data set. At 18 and 21 months, there were several items that were produced by fewer than 5 participants. We, therefore, ran our analyses twice, once including (1) and once excluding (5) these items.

Appendix B:. Examining fit of Spline-IRT model at earliest time points

In Study 1, the spline IRT model produced notably lower DETECT indices than the other two models, particularly for the 16-month data set. To understand this pattern, we extracted the composite scores estimated from 10 of such models and plotted these against total vocabulary and grammar scores (Fig. B1). This allows us to compare the model's estimate of the participant's combined vocabulary and grammar score to the raw scales of each section.

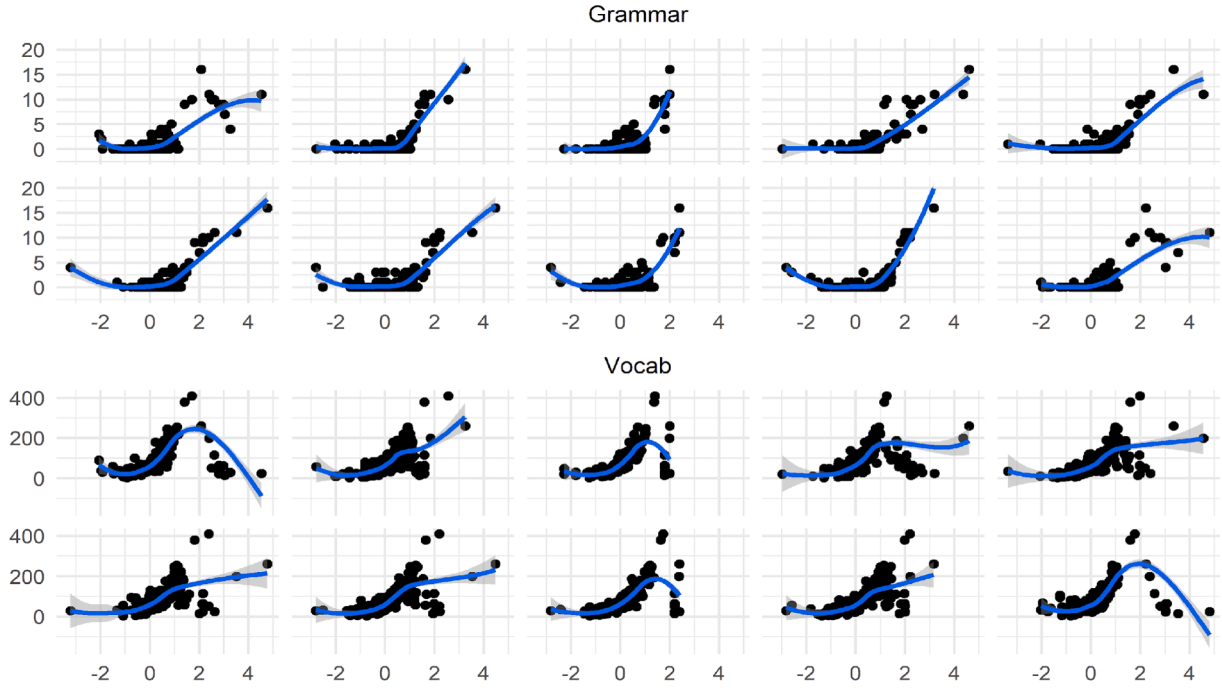

Fig. B1. Plots comparing composite scores estimated in the spline IRT models to raw scores from the vocabulary and grammatical complexity subscales. The X axis indicates the estimated participant-level ability scores. The Y axis indicates the raw score for the relevant section.

As can be seen in this figure the relationship between participant-level ability scores and measured scores (on both grammar and vocabulary) was often non-monotonic. For example, for many datasets ability scores related to vocabulary in an inverse-U shaped curve, with the children with the estimated highest ability scores having quite low vocabularies. As this relationship is apriori implausible we suspect it reflects overfitting. This seems likely given that (a) the spline model is less constrained than the models which specify a link function (the functional form of the item response function is learned from the data rather than specified apriori) (b) the 16-month dataset contained a relatively small number of participants combining words and (c) many items had very low scores. This pattern of over-fitting is particularly problematic because DETECT's primary assumption is that the item-response function is monotone increasing. If the estimates of participant ability relate non-monotonically to the true ability scores, DETECT may fail to identify multidimensionality. See below for a similar plot for data sets including only items that at least 5 participants produced (Fig. B2).

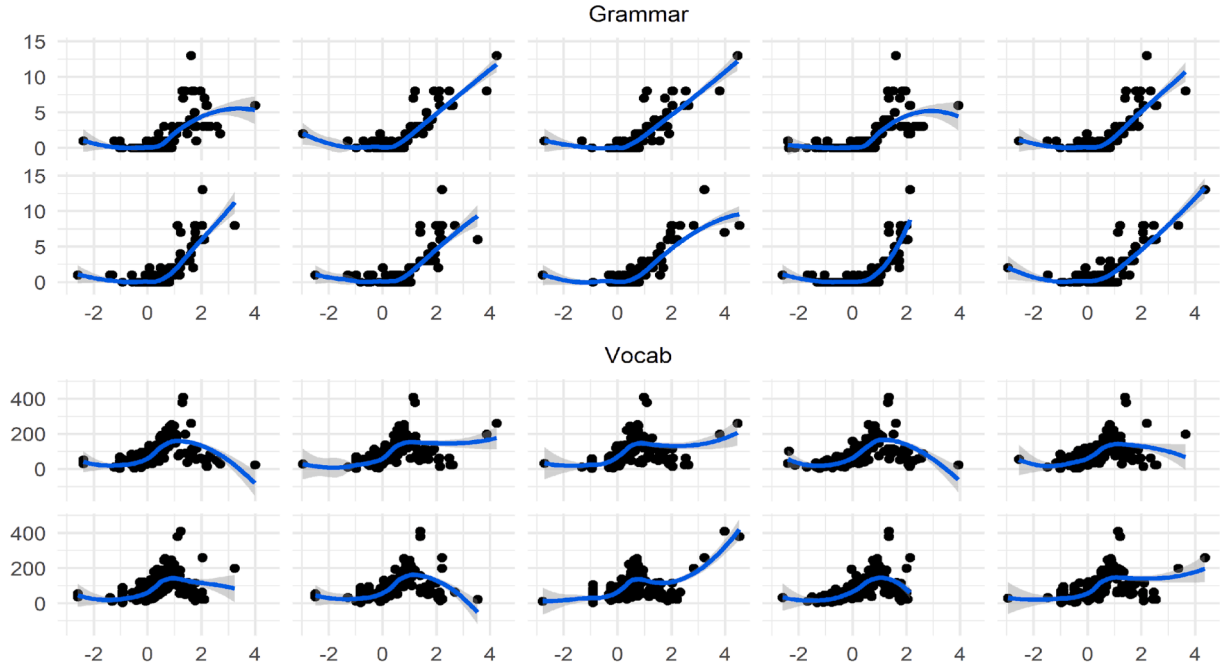

Fig. B2. Plots comparing composite scores estimated in the spline IRT models to raw scores from the vocabulary and grammatical complexity subscales. The X axis indicates the estimated participant-level ability scores. The Y axis indicates the raw score for the relevant section.

Appendix C:. List of items matched across two vocabulary instruments

CLCL	L05
Crocodile	Alligator
Kitty	Kitten
Rooster	Cockerel
Airplane	Aeroplane
Bicycle	Bicycle/bike
Fire Truck	Fire Engine
Motorcycle	Motorbike
Sled	Sleigh
Pusher	Pushchair
Bubbles	Bubble
Lollies	Sweets
Carrots	Carrot
Cheerios	Cornflakes
Donut	Doughnut
French fries	Chips
Gum	Chewing gum
Lollipop	Lollies
Icy pole	Ice lolly
Potato chip	Crisps
Soda/pop	Pop
Spaghetti	Pasta
Boots	Boot
Pants	Trousers
Runners	Trainers
Underpants	Pants/underpants
Zipper	Zip
Buttocks/bottom	Bottom
Feet	Foot
Garbage	Bin
Glasses	Glasses
Serviettes	Napkins
Tissue/kleenex	tissue
Basement	Cellar
Sandbox	Sandpit
Footpath	Path
Country	Countryside
Movie	Cinema
Daddy	Dad
Grandpa	Grandad
Bye	Bye bye
Ring (on phone)	Call (on phone)
Sh.shush.hush	Shh
Hug	cuddle

Data availability

All data, analysis scripts and results in html format, from Study 1 and Study 2, are available on the open-science framework: https://osf.io/4ng2v/?view_only=dc41a530c0fe4a009b57e41e8b90a0ce.

References

Abu-Zhaya, A., & I., & Borovsky, A.. (2022). Do children use multi-word information in real-time sentence comprehension? *Cognitive Science, 46*(3). https://doi.org/10.1111/cogs.13111

Ambridge, B. (2020a). Against stored abstractions: A radical exemplar model of language acquisition. *First Language, 40*(5–6), 509–559. https://doi.org/10.1177/0142723719869731

Ambridge, B. (2020b). Abstractions made of exemplars or "You're all right, and I've changed my mind': Response to commentators. *First Language, 40*(5–6), 640–659. https://doi.org/10.1177/0142723720949723

Ambridge, B., Rowland, C. F., & Pine, J. M. (2008). Is Structure dependence an innate constraint? New experimental evidence from children's complex-question production. *Cognitive Science, 32*(1), 222–255. https://doi.org/10.1080/03640210701703766

Arnon, I., McCauley, S. M., & Christiansen, M. H. (2017). Digging up the building blocks of language: Age-of-acquisition effect for multiword phrases. *Journal of Memory and Language, 92*, 265–280. https://doi.org/10.1016/j.jml.2016.07.004

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*, 67–82. https://doi.org/10.1016/j.jml.2009.09.005

Aslin, R. N., & Newport, E. L. (2012). Statistical Learning: From Acquiring Specific Items to Forming General Rules. *Current Directions in Psychological Science, 21*(3), 170–176. https://doi.org/10.1177/0963721412436806

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. Proceedings of the National Academy of Sciences - PNAS, 106(41), 17284–17289. DOI: 10.1073/pnas.0905638106.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science, 19*, 241–248. https://doi.org/10.1111/j.1467-9280.2008.02075.x

Bates, E., & Goodman, J. (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

Bates, E., Bretherton, I., & Snyder, L. S. (1988). *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge University Press.

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. (2015). When Are Multidimensional Data Unidimensional Enough for Structural Equation Modeling? An Evaluation of the DETECT Multidimensionality Index. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(4), 504–516. https://doi.org/10.1080/10705511.2014.938596

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind, 3*, 52–67. https://doi.org/10.1162/opmi_a_00026

Bresnan, J. (2001). *Lexical-Functional Syntax*. Blackwell.

Brinchmann, E. I., Braeken, J., & Lyster, S. H. (2019). Is there a direct relation between the development of vocabulary and grammar? *Developmental Science, 22*(1). https://doi.org/10.1111/desc.12709

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*(5), 425–455. https://doi.org/10.1080/01690969508407111

Cao, A., & Lewis, M. (2022). Quantifying the syntactic bootstrapping effect in verb learning: A meta-analytic synthesis. *Developmental science, 25*(2). https://doi.org/10.1111/desc.13176

Cattani, A., Floccia, C., Kidd, E., Pettenati, P., Onofrio, D., & Volterra, V. (2019). Gestures and words in naming: Evidence from crosslinguistic and crosscultural comparison. *Language Learning, 69*(3), 709–746. https://doi.org/10.1111/lang.12346

Cazden, C. (1968). The acquisition of noun and verb inflections. *Child Development, 39*, 433–448.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*(2), 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Chomsky, N. (1995). *The minimalist program*. The MIT Press.

Creaghe, N., Quinn, S., & Kidd, E. (2021). Symbolic play provides a fertile context for language development. *Infancy, 26*(6), 980–1010. https://doi.org/10.1111/infa.12422

Dabrowska, E. (2004). Rules or schemas? Evidence from Polish. *Language and Cognitive Processes, 19*(2), 225–271. https://doi.org/10.1080/01690960344000170

Dąbrowska, E., & Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics, 16*(3), 437–474. https://doi.org/10.1515/cogl.2005.16.3.437

Dale, P. S., Dionne, G., Eley, T. C., & Plomin, R. (2000). Lexical and grammatical development: A behavioural genetic perspective. *Journal of Child Language, 27*(3), 619–642. https://doi.org/10.1017/S0305000900004281

Day, T. K. M., & Elison, J. T. (2022). A broadened estimate of syntactic and lexical ability from the MB-CDI. *Journal of Child Language, 49*(3), 615–632. https://doi.org/10.1017/S0305000921000283

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 369*(1634). https://doi.org/10.1098/rstb.2012.0394

Dionne, G., Dale, P. S., Boivin, M., & Plomin, R. (2003). Genetic evidence for didirectional effects of early lexical and grammatical development. *Child Development, 74*(2), 394–412. https://doi.org/10.1111/1467-8624.7402005

Dixon, J. A., & Marchman, V. A. (2007). Grammar and the lexicon: Developmental ordering in language acquisition. *Child Development, 78*(1), 190–212. https://doi.org/10.1111/j.1467-8624.2007.00992.x

Dolan, C. V., Oort, F. J., Stoel, R. D., & Wicherts, J. M. (2009). Testing measurement invariance in the target rotates multigroup exploratory factor model. *Structural Equation Modeling, 16*, 295–314. https://doi.org/10.1080/10705510902751416

Donnelly, S., & Kidd, E. (2020). Individual differences in lexical processing efficiency and vocabulary in toddlers: A longitudinal investigation. *Journal of Experimental Child Psychology, 192*, Article 104781. https://doi.org/10.1016/j.jecp.2019.104781

Donnelly, S., & Kidd, E. (2021). The longitudinal relationship netween conversational yurn-taking and vocabulary growth in early language development. *Child Development, 92*(2), 609–625. https://doi.org/10.1111/cdev.13511

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition, 203*. https://doi.org/10.1016/j.cognition.2020.104348

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences - PNAS, 113*(41), E6256–E6262. https://doi.org/10.1073/pnas.1612132113

Fenson, L., Bates, E., Dale, P., Marchman, V., Reznick, J., & Thal, D. (2007). Mac-Arthur Bates communcative development inventories (2nd Edition). Brookes.

Fernald, A., Marchman, V., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months.

Fisher, C. (2002). The Role of Abstract Syntactic Knowledge in Language Acquisition: A Reply to Tomasello (2000). *Cognition, 82*(3), 259–278. https://doi.org/10.1016/S0010-0277(01)00159-7

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language, 44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *The Variability and Consistency in Early Language Learning: The Wordbank Project*. MIT Press.

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science, 17*(8), 684–691. https://doi.org/10.1111/j.1467-9280.2006.01767.x

Gleitman, L. (1990). The Structural Sources of Verb Meanings. *Language Acquisition, 1*(1), 3–55. https://doi.org/10.1207/s15327817la0101_2

Goldberg, A. E. (1995). Constructions: A construction grammar approach to argument structure. In Xi+265pp, Chicago, IL: U Chicago Press, 1995. University of Chicago Press. https://go.exlibris.link/dD7kQQJd.

Hagoort, P. (2013). MUC (Memory, Unification, Control) and beyond. *Frontiers in Psychology, 4*, 416. https://doi.org/10.3389/fpsyg.2013.00416

Hayes, B. K., Dunn, J. C., Joubert, A., & Taylor, R. (2017). Comparing single- and dual-process models of memory development. *Developmental Science, 20*(6). https://doi.org/10.1111/desc.12469

Hoff, E., Quinn, J. M., & Giguere, D. (2018). What explains the correlation between growth in vocabulary and grammar? New evidence from latent change score analyses of simultaneous bilingual development. *Developmental Science, 21*(2). https://doi.org/10.1111/desc.12536

Huang, Q., & Bolt, D. M. (2024). Unipolar IRT and the Author Recognition Test (ART). *Behavior Research Methods, 56*(6), 5406–5423. https://doi.org/10.3758/s13428-023-02275-2

Jackendoff, R. (2013) Constructions in the Parallel Architecture. In Hoffman, T. & Trousdale, G. (Eds.), The Oxford Handbook of Construction Grammar. (pp 70 – 92). Oxford. (DOI: 10.1093/oxfordhb/9780195396683.013.0005.

Jang, E. E., & Roussos, L. (2007). An Investigation into the Dimensionality of TOEFL Using Conditional Covariance-Based Nonparametric Approach. *Journal of Educational Measurement, 44*(1), 1–21. https://doi.org/10.1111/j.1745-3984.2007.00024.x

Jolsvai, H., McCauley, S. M., & Christiansen, M. H. (2020). Meaningfulness beats frequency in multiword chunk processing. *Cognitive science, 44*(10). https://doi.org/10.1111/cogs.12885

Kidd, E., & Donnelly, S. (2020). Individual differences in first language acquisition. *Annual Review of Linguistics, 6*(1), 319–340. https://doi.org/10.1146/annurev-linguistics-011619-030326

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences, 22*(2), 154–169. https://doi.org/10.1016/j.tics.2017.11.006

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language, 42*(6), 703–735. https://doi.org/10.1177/01427237211066405

Kidd, E., Junge, C. M., Spokes, T., Morrison, L., & Cutler, A. (2018). Individual differences in infant speech segmentation: Achieving the lexical shift. *Infancy, 23*(6), 770–794. https://doi.org/10.1111/infa.12256

Kievit, R. A., Lindenberger, U., Goodyer, I. M., Jones, P. B., Fonagy, P., Bullmore, E. T., & Dolan, R. J. (2017). Mutualistic Coupling Between Vocabulary and Reasoning Supports Cognitive Development During Late Adolescence and Early Adulthood. *Psychological Science, 28*(10), 1419–1431. https://doi.org/10.1177/0956797617710785

Kievit, R. A., Hofman, A. D., Nation, K., et al. (2019). Mutualistic Coupling Between Vocabulary and Reasoning in Young Children: A Replication and Extension of the Study by Kievit et al. (2017). *Psychological Science, 30*(8), 1245–1252. https://doi.org/10.1177/0956797619841265

Kim, H. (1994). New techniques for the dimensionality assessment of standardized test data (Order No. 9512427). Available from ProQuest One Academic. (304133056).

Koller, I., & Alexandrowicz, R. W. (2010). Eine psychometrische Analyse der ZAREKI-R mittels Rasch-Modellen [A psychometric analysis of the ZAREKI-R using Rasch-models]. *Diagnostica, 56*, 57–67.

Kumarage, S., Donnelly, S., & Kidd, E. (2022). Implicit learning of structure across time: A longitudinal investigation of syntactic priming in young English-acquiring children. *Journal of Memory and Language, 127*. https://doi.org/10.1016/j.jml.2022.104374

Language and Reading Research Consortium. (2015). The dimensionality of language ability in young children. *Child Development, 86*(6), 1948–1965. https://doi.org/10.1111/cdev.12450

Lieven, E., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language, 24*(1), 187–219. https://doi.org/10.1017/S0305000996002930

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition, 6*, 312–319. https://doi.org/10.3758/BF03197461

Mair, P. (2020). Modern Psychometrics with R. *R Package*. https://doi.org/10.32614/CRAN.package.MPsychoR

Marchman, V. A., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language, 21*(2), 339–366. https://doi.org/10.1017/S0305000900009302

Marchman, V. A., Martinez-Sussmann, C., & Dale, P. S. (2004). The language-specific nature of grammatical development: Evidence from bilingual language learners. *Developmental Science, 7*(2), 212–224. https://doi.org/10.1111/j.1467-7687.2004.00340.x

Marchman, V. A., Dale, P. S., & Fenson, L. (2023). *The MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual* (3rd Edn.). Baltimore, MD: Brookes Publishing Co.

Marcus, G. F. (1992). Overregularization in Language Acquisition. *Monographs of the Society for Research in Child Development, 57*(4), 1–182. https://doi.org/10.1111/j.1540-5834.1992.tb00313.x

Mayor, J., & Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from CDI analysis: A statistical estimate of vocabulary size. *Developmental Science, 14*(4), 769–785. https://doi.org/10.1111/j.1467-7687.2010.01024.x

McCauley, S., Bannard, C., Theakston, A., Davis, M., Cameron-Faulkner, T., & Ambridge, B. (2021). Multiword units lead to errors of commission in children's spontaneous production: "What corpus data can tell us?". *Developmental Science, 24*(6). https://doi.org/10.1111/desc.13125

McCauley, & Christiansen, M. H.. (2019a). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review, 126*(1), 1–51. https://doi.org/10.1037/rev0000126

McCauley, & Christiansen, M. H.. (2019b). Modeling Children's early linguistic productivity through the automatic discovery and use of lexically-based frames. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society, Cognitive Science Society* (pp. 782–788).

Meints, K., Fletcher, K., & Just, J. (2017). The Lincoln toddler communicative development inventory—A UK adaptation of the MacArthur-Bates communicative development inventory: Words and sentences (Toddler form).

Messenger, K., & Fisher, C. (2018). Mistakes weren't made: Three-year-olds' comprehension of novel-verb passives provides evidence for early abstract syntax. *Cognition, 178*, 118–132. https://doi.org/10.1016/j.cognition.2018.05.002

Meylan, S. C., Frank, M. C., Roy, B. C., & Levy, R. (2017). The emergence of an abstract grammatical category in children's early speech. *Psychological Science, 28*(2), 181–192. https://doi.org/10.1177/0956797616677753

Monaghan, P., Donnelly, S., Alcock, K., Bidgood, A., Cain, K., Durrant, S., … Rowland, C. F. (2023). Learning to generalise but not segment an artificial language at 17 months predicts children's language skills 3 years later. *Cognitive Psychology, 147*, 101607–101607. https://doi.org/10.1016/j.cogpsych.2023.101607.

Naigles, L. R. (2002). Form is easy, meaning is hard: Resolving a paradox in early child language. *Cognition, 86*(2), 157–199. https://doi.org/10.1016/S0010-0277(02)00177-4

Nusbaum, H. C., & Goodman, J. C. (1994). Learning to hear speech as spoken language. *The Development of Speech Perception: The Transition from Speech Sounds to Spoekn Words* (Vol. 1–Book).

Pérez-Leroux, A. T., Castilla-Earls, A. P., & Brunner, J. (2012). General and Specific Effects of Lexicon in Grammar: Determiner and Object Pronoun Omissions in Child Spanish. *Journal of Speech, Language, and Hearing Research, 55*(2), 313–327. https://doi.org/10.1044/1092-4388(2011/10-0004)

Peter, M. S., Durrant, S., Jessop, A., Bidgood, A., Pine, J. M., & Rowland, C. F. (2019). Does speed of processing or vocabulary size predict later language growth in toddlers? *Cognitive Psychology, 115*, Article 101238. https://doi.org/10.1016/j.cogpsych.2019.101238

Pinker, S. (1999). Words and rules: The ingredients of language. *Weidenfeld & Nicolson*.

Pinker, S. (1994). *The language instinct* ((1st ed.). W.). Morrow and Co.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition, 48*(1), 21–69. https://doi.org/10.1016/0010-0277(93)90057-3

Pollard, C. J., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. Center for the Study of Language and Information.

Reckase, M. D. (2009). Multidimensional Item Response Theory. *Springer*. https://doi.org/10.1016/0010-0277(93)90057-3

Reilly, S., Wake, M., Bavin, E. L., Prior, M., Williams, J., Bretherton, L., … Ukoumunne, O. C. (2007). Predicting language at 2 years of age: a prospective community study. *Pediatrics, 120*(6), e1441–e1449. https://doi.org/10.1542/peds.2007-0045

Robitzsch, A. (2022). sirt: Supplementary Item Response Theory Models. https://CRAN.R-project.org/package=sirt.

Rowland, C., & Fletcher, S. L. (2006). The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language, 33*(4), 859–877. https://doi.org/10.1017/S0305000906007537

Rowland, C. F., Chang, F., Ambridge, B., Pine, J. M., & Lieven, E. V. M. (2012). The development of abstract syntax: Evidence from structural priming and the lexical boost. *Cognition, 125*(1), 49–63. https://doi.org/10.1016/j.cognition.2012.06.008

Rowland, C. F., Bidgood, A., Durrant, S., Peter, M., Pine, J. M., & Jago, L. (2018). The language 0-5 project. *Unpublished manuscript, University of Liverpool*. https://doi.org/10.17605/OSF.IO/KAU5F.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences - PNAS, 112*(41), 12663–12668. https://doi.org/10.1073/pnas.1419773112

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science (American Association for the Advancement of Science), 274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Shain, C., Kean, H., Lipkin, B., Affourtit, J., Siegelman, M., Mollica, F., & Federenko, E. (2020). "Constituent length" effects in fMRI do not provide evidence for abstract sytnactic processing. *BioRxiv*.

Sijtsma, K., & Meijer, R. R. (2006). Nonparametric Item Response Theory and Special Topics. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26). Psychometrics. Elsevier B.V. https://doi.org/10.1016/S0169-7161(06)26022-X.

Snider, N., & Arnon, I. (2012). A unified lexicon and grammar? Compositional and non-compositional phrases in the lexicon. In S. Gries, & D. Divjak (Eds.), *Frequency effects in language*. Berlin: Mouton deGruyter.

Steedman, M. (2000). *The syntactic process*. MIT Press.

Stephens, R. G., Matzke, D., & Hayes, B. K. (2019). Disappearing dissociations in experimental psychology: Using state-trace analysis to test for multiple processes. *Journal of Mathematical Psychology, 90*, 3–22. https://doi.org/10.1016/j.jmp.2018.11.003

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*(4), 331–354. https://doi.org/10.1177/014662169602000403

Thal, D. J., Marchman, V. A., & Tomblin, J. B. (2013). Late-talking toddlers: Characterization and prediction of continued delay. In *Late talkers: Language development, interventions, and outcomes* (pp. 169–201). Paul H. Brookes Publishing Co.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition, 74*(3), 209–253. https://doi.org/10.1016/S0010-0277(99)00069-4

Tomblin, J. B., & Zhang, X. (2006). The Dimensionality of language ability in school-age children. *Journal of Speech, Language, and Hearing Research, 49*(6), 1193–1208. https://doi.org/10.1044/1092-4388(2006/086)

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition, 92*(1), 231–270. https://doi.org/10.1016/j.cognition.2003.10.008

Ullman, M. T., Clark, G. M., Pullman, M. Y., Lovelett, J. T., Pierpont, E. I., Jiang, X., & Turkeltaub, P. E. (2024). The neuroanatomy of developmental language disorder: A systematic review and meta-analysis. *Nature Human Behaviour, 8*(5), 962–975. https://doi.org/10.1038/s41562-024-01843-6

Valentini, A., & Serratrice, L. (2021). What can bilingual children tell us about the developmental relationship between vocabulary and grammar? *Cognitive Science, 45*(11). https://doi.org/10.1111/cogs.13062

Van Der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition, 40*(2), 145–160. https://doi.org/10.3758/s13421-011-0158-0

Zhang, J. (2013). A procedure for dimensionality analyses of response data from various test designs. *Psychometrika, 78*(1), 37–58. https://doi.org/10.1007/s11336-012-9287-z

Zhang, J., & Stout, W. (1999). The Theoretical DETECT Index of Dimensionality and Its Application to Approximate Simple Structure. *Psychometrika, 64*(2), 213–249. https://doi.org/10.1007/bf02294536