HHS Public Access

Author manuscript

*Int J Lang Commun Disord*. Author manuscript; available in PMC 2019 August 06.

Published in final edited form as:

*Int J Lang Commun Disord*. 2019 January ; 54(1): 79–94. doi:10.1111/1460-6984.12433.

# Reconceptualizing developmental language disorder as a spectrum disorder: issues and evidence

**Hope S. Lancaster**[†], **Stephen Camarata**[‡]

[†]Department of Speech and Hearing Science, Arizona State University, Tempe, AZ, USA

[‡]Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, TN, USA

## Abstract

**Background:** There is considerable variability in the presentation of developmental language disorder (DLD). Disagreement amongst professionals about how to characterize and interpret the variability complicates both the research on understanding the nature of DLD and the best clinical framework for diagnosing and treating children with DLD. We describe and statistically examine three primary possible models for characterizing the variability in presentation in DLD: predictable subtypes; individual differences; and continuum/spectrum.

**Aims:** To test these three models of DLD in a population-based sample using two distinct types of cluster analyses.

**Methods & Procedures:** This study included children with DLD ($n = 505$) from the US Epidemiological Study of Language Impairment database. All available language and cognitive measures were included. Two cluster methods were used: Ward's method and *K*-means. Optimal cluster sizes were selected using Bayesian information criteria (BIC). Bootstrapping and permutation methods were used to evaluate randomness of clustering.

**Outcomes & Results:** Both clustering analyses yielded more than 10 clusters, and the clusters did not have spatial distinction: many of these clusters were not clinically interpretable. However, tests of random clustering revealed that the cluster solutions obtained did *not* arise from random aggregation.

**Conclusions & Implications:** Non-random clustering coupled with a large number of non-interpretable subtypes provides empirical support for the continuum/spectrum and individual differences models. Although there was substantial support for the continuum/spectrum model and weaker support for the individual differences model, additional research testing these models should be completed. Based on these results, clinicians working with children with DLD should focus on creating treatment plans that address the severity of functioning rather than seeking to

Address correspondence to: Hope Lancaster, Speech and Hearing Science, Arizona State University, PO Box 85281, Tempe, AZ 85281-0102, USA; hope.lancaster@asu.edu.

***Declaration of interest:*** The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

identify and treat distinct subtypes. Additional consideration should be given to reconceptualizing DLD as a spectrum condition.

**Keywords**

developmental language impairment; quantitative; specific language impairment

## Introduction

Developmental language disorder (DLD) occurs when a child has severe and persistent difficulties with language understanding and/or functioning that impact daily life and/or educational attainment when no other medical condition is present (i.e., hearing loss) (Bishop *et al.* 2017). It is a complex phenomenon, characterized by high levels of variability in presentation. Previously, the field has adopted and abandoned divergent diagnostic subtype categories, such as expressive-only and mixed receptive-expressive (Diagnostic and Statistical Manual of Mental Disorders, 4th Edition—DSM-IV, American Psychiatric Association, 1994) to address the variability in clinical presentation. Some professionals have argued for diagnostic classification based on primary domain of language deficits (e.g., grammatical specific language impairment—SLI; Bishop 2004) and others have proposed using severity of functioning as an explanation of variability (e.g., mild, moderate or severe; Leonard 2009, 2010).

Moreover, disagreement about how to explain and characterize this variability has made it difficult for the field to reach consensus about inclusionary criteria and terminology. For example, in a recent report on the CATALISE study, Bishop *et al.* (2017) point to the adverse impact of this problem, arguing that: ‘lack of agreement about criteria and terminology for children’s language problems affects access to services as well as hindering research and practice’ (1). The CATALISE study, which is focused on creating consensus among professionals about the criteria and terminology for DLD, illuminated key agreements and disagreements about criteria and terminology. Therefore, there is a compelling need for a data-driven exploration to investigate approaches to best capture and classify the persistent variability in DLD presentation (Bishop *et al.* 2016, 2017, Bishop 2017). There are several potential sources for this variability, ranging from demographic factors (e.g., socioeconomic status of child) to co-morbidities to lack of agreement about the underlying theoretical structure of the DLD population. The discipline requires a clearer, empirically tested model in a population sample of DLD to winnow competing theoretical models and guide clinical practice (Raghavan *et al.* 2018).

Another ongoing source of variability is differing use of nonverbal or ‘performance intelligence’ information (Bishop 2017) for classification. For example, Bishop (2017) states that respondents to the CATALISE study disagreed about whether performance IQ should be employed to determine whether a child has a language disorder. Historically, performance IQ has been employed to distinguish DLD from global intellectual disability (Camarata and Swisher 1990, Stark and Tallal 1981) and many studies of DLD have included measures of performance IQ. However, others have argued that DLD can (and should) also include children who fall above the IQ cut-off for broader intellectual disability but below a 1 SD

(standard deviation) cut-off in addition to all those falling above the 1 SD level (e.g. Tomblin *et al.* 1997). As Bishop and colleagues compellingly argue, there is a pressing need to clarify the nature of DLD at a foundational level. From a broad perspective, population studies are needed to test whether the condition itself represents an aggregate of predictable (and clinically relevant) subtypes, an amalgamation of individual differences with individual variation in specific clinical features (Leonard 2014), or, akin to autism spectrum disorder (ASD), a spectrum disorder across severity levels rather than distinct subtypes or individual difference.

Historically, there have been two primary models used to explain the underlying population structure for children with DLD: (1) subtypes (Bishop 2004, Bishop *et al.* 2000a, 2000b, Conti-Ramsden *et al.* 1997); and (2) individual differences (Leonard 2014). These two models have been tested extensively and have yet to produce agreement in the field about which is a better fit with the population (Bishop *et al.* 2017). The focus of past research exploring these models has been to confirm a theoretical approach about what subtypes are present (e.g., Bishop *et al.* 2000a) or to demonstrate the utility of psychology measures for subtyping children with DLD (Conti-Ramsden *et al.* 1997). However, few studies have directly compared subtype and the individual differences models using a single analytical strategy within a population sample, which is an important element in classification of a clinical typology (Raghavan *et al.* 2018).

It is noteworthy that this is neither a new problem nor is it unique to DLD: researchers describing other neurodevelopmental disorders (e.g., autism, attention-deficit hyperactivity disorder—ADHD), have similarly struggled with how to explain variability in presentation as clinical understanding of the disorder grew (as an example for these conditions, see the successive revisions of DSM-III, IV and V). For example, there is a venerable history of research and clinical models of autism wherein subtypes and individual difference models (Kanner 1968) were hotly debated and scrutinized in the past, but researchers and clinicians ultimately have discarded these and adopted a third model, a continuum/spectrum model for the condition.

This was ultimately adopted only after a priori subtype and individual difference models were not confirmed statistically in population samples or clinically in terms of accurate differential diagnosis and treatment outcomes (Wing 1997). To our knowledge, research on DLD has not explored a continuum/spectrum model, but given the ongoing difficulty in capturing DLD that parallels what was seen in other neurodevelopmental disabilities such as ASD and ADHD, a spectrum approach merits consideration. If validated, a spectrum conceptualization of DLD could potentially have important theoretical and clinical impacts on the field's understanding of aetiology, inclusionary criteria, diagnosis, assessment and treatment of this condition. In addition to providing theoretical arguments supporting or criticizing these different views of DLD, it is noteworthy that the validity of these models can be tested. Therefore, we will discuss the merits of each model, how each model handles performance IQ, and present the rationale and criteria for testing these models within a cluster analysis framework.

Subtype modelling is predicated on the testable hypothesis that the variability in children with DLD is the result of an aggregation of multiple discrete (predictable and replicable) groups. A key assumption in a subtype model is that children with DLD make up a distinct population compared with typical language learners and that this distinct population can be statistically resolved into predicable typologies based on shared features in the subtypes. This is akin to previous and now largely abandoned efforts to subtype autism into discrete groups of Pervasive Developmental Disorder (PDD) such as PDD-Asperger Syndrome, PDD-Autism, PDD-Childhood Disintegrative Disorder and PDD-Not Otherwise Specified (NOS). For example, Prior *et al.* (1998) directly tested for DSM-IV designated subgroups of autism and concluded:

> Although subgroups were identified which bore some relationship to clinical differentiation of autistic, Asperger syndrome, and Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS) cases, the nature of the differences between them appeared strongly related to ability variables. Examination of the kinds of behaviours that differentiated the groups suggested that a spectrum of autistic disorders on which children differ primarily in term of degrees of social and cognitive impairments could explain the findings. (893)

These types of results, wherein the prevailing subtypes of autism could not be reliably sorted in population samples directly lead to abandoning subtypes (Lord and Risi 1998), culminating in the codification of autism as a spectrum disorder in the most recent edition of the DSM (Frazier *et al.* 2012).

A subtype model suggests that the majority of the variability in any condition (such as DLD) is *explanatory* and can be used to create *predictable*, relatively homogenous subgroups (e.g., Type 1 and Type 2 diabetes mellitus). Conceptually, this means that a general variability can be parsed into defined regions of relative stability so the overall population-wise variability arises from variation between these subgroups.

As an example of this approach in DLD, Tomblin *et al.* (1997) used performance IQ scores to create two subtypes of children with language problems, they designated as *specific* (SLI) and non-specific language impairment (NLI) in the Epidemiological Study of Specific Language Impairment (EpiSLI) database. Similarly, Bishop's (2004) subtypes include 'typical' SLI (formerly phonological-syntactic), severe receptive language disorder (formerly verbal auditory agnosia), developmental verbal dyspraxia, and pragmatic language disorder. Furthermore, validation of subtype models also requires that we can create assessments and interventions that statistically (and clinically) capture each specific subtype of DLD. Analytically, a subtype model is confirmed when the outcome of statistical cluster analyses meets *both* of the following criteria: (1) yield a small number of clusters based on similar features; and (2) clusters are not random.

Because subtyping has an intuitive appeal, there has been a long history of attempts to create clinically valid subtypes of DLD. A detailed review of the extensive literature on DLD subtypes is beyond the scope of this paper, but selected examples demonstrate the parameters of this perspective. Subtypes of DLD have been conceptualized by separating subtypes on clinical measures (Rapin and Allen 1987), psychometric tests (Aram and Nation

1975) and linguistic domains in addition to performance IQ levels (Tomblin *et al.* 1997). Past researchers have used a variety of statistical methods including: factor analysis (Aram and Nation 1975), cluster analysis (Conti-Ramsden and Botting, 1999, Conti-Ramsden *et al.* 1997, Tomblin and Zhang 1999), qualitative-descriptive accounts (Rapin and Allen 1987), and standardized or experimental measures (Bishop 2004, Bishop *et al.* 2000a, 2000b, Bishop and Rosenbloom 1987) as evidence to support each subtype scheme. Although some researchers have reported support for potential subtypes, either theoretically driven (Bishop and Rosenbloom 1987, Bishop *et al.* 2000a, 2000b, Bishop 2004) or empirically driven (Conti-Ramsden *et al.* 1997, Conti-Ramsden and Botting 1999), and despite the promising initial results for some of these subtyping approaches, it is fair to say that these efforts have yielded, at best, ambiguous results on the validity and presence of subtypes (Conti-Ramsden *et al.* 1997, Rapin and Allen 1987, Tomblin and Zhang 1999, Botting and Conti-Ramsden 2004, Conti-Ramsden and Botting 1999, Bishop *et al.* 2000a, Bishop and Rosenbloom 1987, Feagans, Porter, Child, and Applebaum 1986). Past research has shown the subtypes to be unstable and not reproducible (Conti-Ramsden and Botting 1999, Tomblin and Zhang 1999) and more recent research has yielded active counter evidence on the existence of subtypes of DLD (Dollaghan 2011, Reilly *et al.* 2014b, Bishop *et al.* 2017).

In contrast to subtyping approaches, individual differences models posit that population variability is the byproduct of each child's unique profile (see the review by Leonard 2014). Individual differences models are predicated on the hypothesis that population variability is random, so that predictable homogeneous groups *cannot be detected*. From a clinical perspective, individual differences modelling implies that clinical approaches to assessment and treatment should be focused on supporting all children with DLD using personalized treatment plans tailored to each individual profile rather than on diagnosing subtypes. Some advocates of the individual differences model argue that children with DLD are not a distinct language learning population from typical language learners, but rather make up the left-hand tail of a normal distribution. An individual differences approach to DLD is supported by past research demonstrating that children with DLD display individual traits that cannot be readily or consistently be organized into subtypes (Leonard 1989, 2009, 2014). A strength of the individual differences model is that it explains the wide variation in performance on various aspects of language (e.g., semantics, morphosyntax, pragmatics) in children with DLD and matches with current clinical practices to create personalized treatment plans based on individual strengths and weaknesses and on functional needs.

However, thus far, individual differences models have been unable to account for consistent language differences between children classified as DLD and typical language learners on multiple, predictable dimensions (e.g., tense-marking errors; Rice *et al.* 1998, 2004), or reduced syntactic complexity (Mackie and Dockrell 2004, Nippold *et al.* 2009). That is, there do appear to be relatively few shared features around morphosyntax that are inconsistent with subtyping, but also inconsistent with random variation as would be expected if these arose from individual differences. Statistical testing and validation of an individual differences model will result in both: (1) a cluster analysis that yields a very large number of clusters; and (2) clusters that are formed randomly.

A third model, heretofore untested in the literature on DLD, namely a continuum/spectrum model, has emerged as a useful (and valid) model in several other neurodevelopmental disorder populations such as learning disability (LD), ADHD (Mayes *et al.* 2000) and autism (ASD; Frazier *et al.* 2012), which were initially viewed as subtypes and/or individual differences. A continuum/spectrum model builds on elements of both subtype and individual differences models. Namely, (1) that children with DLD are a distinct group of language learners; and (2) that there is a need to explain and clinically address individual strengths and needs. In contrast to competing models, a continuum/spectrum model does not presume subtypes, but does posit predictable (non-random) performance on shared features of the condition (e.g., a general weakness in morphosyntax). Also, a continuum/spectrum model addresses the performance IQ issue by positing that performance IQ is simply one dimension of functioning that children with DLD can vary on rather than assuming convergence or divergence with typically developing children.

Past research on the subtype and individual differences models provides initial indirect support for a continuum/spectrum approach. The continuum/spectrum model combines aspects of the subtype and individual differences models by hypothesizing that there is an underlying commonality among children with DLD, but that each child will have a slightly different presentation within the broader spectrum condition. Specifically, a continuum/spectrum model includes key shared features for all members of the posited clinical typology (e.g., DLD), but includes the assumption that there are no consistently identifiable profiles shared by homogeneous subgroups within the broader clinical population.

A potential advantage of a continuum/spectrum model is that it can explain the consistent differences between children with DLD and typical language learners while at the same time explaining why some children with DLD are, comparatively, better at one aspect of language than their other peers with DLD. Moreover, testing the continuum/spectrum model does not require a matched sample of typical language learners once the presence of the clinical feature(s) has been established because a continuum/spectrum model is an internal analysis for the clinical population. A spectrum model, does however, require that the clinical group be considered a unique population that can be reliably and validly distinguished from typically developing children.

Evidence for the continuum/spectrum model has both theoretical and clinical implications. Theoretically, if confirmed, this would indicate a need for additional research into identification of the core elements of DLD wherein severity features rather than subtype features are assayed and analyzed. Clinically, a continuum/spectrum model would support developing individualized treatment plans to address the core features of the condition as expressed in a specific patient and raise a demand for development of new assessment procedures focused on core elements and functional capacity for individuals. When using cluster analysis to statistically test for a continuum/spectrum model, results should yield (1) clusters that are statistically salient which are not based on any interpretable (a priori) features; and (2) that these clusters are not randomly formed.

Cluster analysis is a type of ‘unsupervised’ machine learning, wherein unclassified data are analyzed to determine whether there are probabilistic patterns detected in the data set. This

means that the analysis is not actively guided in any way. This approach is ideal when researchers are interested in identifying what natural groupings occur in a data set, in contrast to a supervised approach which uses predefined groupings and active guidance from the researchers. For this paper, cluster analysis is the best choice because there are several competing subtype models, many of which have more than two groups which would be too complicated for supervised methods. Additionally, we wanted to compare the theoretical models of subtypes, individual differences and continuum/spectrum with each other, which is not possible using supervised methods wherein the researchers preselect features such as different parameter weights (e.g., for IQ scores). Therefore, we tested the three models using cluster analysis because cluster analysis: (1) works well with a small number of variables; (2) focuses on the case level, thus allowing us to focus on the underlying structure of a population; (3) can test for the presence of more than three groups; and 4) does not require the use of predefined groupings.

Cluster analysis has been used in multiple populations to test for subtypes (e.g., Zaihra *et al.* 2016) including DLD (Conti-Ramsden *et al.* 1997, Tomblin and Zhang 1999) or continuous/spectrum population structures (e.g., Wiggins *et al.* 2012). Based on the underlying theory of each model, we can use two aspects to distinguish the models from each other: number of clusters and degree of randomness when clustering. Degree of randomness refers to how likely the clusters obtained are random versus true representation of the relationship between cases. Using these two aspects we can place each model in a quadrant as seen in table 1.

The purpose of this study is to test whether any of the three models best explains the variability in a population-based sample of DLD using cluster analysis. This study adds to the current DLD literature by providing a direct comparison between three models of DLD (i.e., subtypes, individual differences, continuum/spectrum) using a single omnibus analysis, which is not available in the literature currently. This studies differs from and fills in gaps left by past research in three ways: (1) it includes unsupervised inclusion of performance IQ measures in the cluster analysis; (2) it does not use predefined groupings; and (3) it compares three competing theoretical models using the same unbiased statistical approach. The study is one of the first to investigate a continuum/spectrum model using an epidemiological sample. The research questions are:

- Does a subtype model best explain the variability in a population-based sample of children with DLD?
- Does an individual differences model best explain the variability a population-based sample of children with DLD?
- Does a continuum/spectrum model best explain the variability a population-based sample of children with DLD?

## Materials and methods

The Vanderbilt University institutional review board approved this study.

## Data

This study uses the Epidemiological Study of Specific Language Impairment (EpiSLI; Tomblin 2010) kindergarten database. The database is an epidemiological sample of children with DLD and a random selection of a sample of peers without DLD collected at a *single time point*. The original purpose was to estimate the true occurrence rate of DLD in the United States in kindergarten.

## Participants

The EpiSLI kindergarten database includes over 1000 participants ($N = 1920$) with and without DLD. There are 505 cases that Tomblin and colleagues originally classified as either SLI or NLI. We selected cases identified as SLI and NLI for two reasons: (1) because the focus of this analysis was on the underlying population structure of all children diagnosed with DLD, not on the underlying structure of language use; and (2) the ratio to case control (505:1226) is high enough that including typically developing children would potentially obscure any clustering within the DLD group. This selection process resulted in the DLD data set herein. Table 2 provides descriptive information about the participants. Tomblin *et al.* (1996) required that children have two language composite (e.g., vocabulary, grammar or narrative) *z*-scores $< -1.25$. This diagnostic practice was validated against clinical judgement and Tomblin *et al.* (1996) found high concordance with clinical judgement, as well as high sensitivity and specificity when using –1.25 *z*-scores over other options. Although only 29% of children identified by Tomblin and colleagues were receiving language therapy (Tomblin *et al.* 1997), all children were significantly different from non-DLD children and represent not just a clinical sampling of DLD but all children with DLD. The grand mean *z*-score for language ability was –1.62 (SD = 0.44) and the mean for performance IQ *z*-sores was –0.51 (SD = 0.86).

## Variables

This study included all the possible language and performance IQ variables from the EpiSLI database. There were eight language measures and two performance IQ measures (Tomblin *et al.* 1997). These variables have been described extensively in Tomblin's work (for a review, see Tomblin *et al.* 1996, 1997), but we provide a brief description of the variables below.

Six of the language variables were derived from the Test of Language Development—Primary 2nd Edition (TOLD-P:2) (Newcomer and Hammill 1988); these variables were: (1) Picture Vocabulary, (2) Oral Vocabulary, (3) Grammatical Understanding, (4) Grammatical Completion, (5) Sentence Imitation and (6) Word Articulation. Two language variables were derived from a narrative task Culatta *et al.* (1983) developed previously: (1) narrative recall and (2) narrative comprehension. Tomblin employed the TOLD-P:2 (Newcomer and Hammill 1988) and two narrative tasks from Culatta *et al.* (1983) to measure receptive and expressive abilities for vocabulary, grammar and narrative (Tomblin *et al.* 1996). Vocabulary and grammar were targeted because these are the most common domains for deficits in children with language impairment (Bishop 1997, Leonard 2014, Newcomer and Hammill 1988). Narrative tasks were included because narrative performance predicts academic performance above and beyond other language abilities (Feagans *et al.* 1986, Tomblin *et al.*

1996). Item analysis validated the TOLD-P:2 and this instrument is viewed as psychometrically sound. Newcomer and Hammill (1988) provide discriminative power and difficulty level for each subtest, with all subtests at or above criteria, thus indicating that the subtests are valid. Newcomer and Hammill also provide data about the reliability of each subtest by investigating the internal consistency and stability. The data provided indicate that overall the subtests are reliable therefore any lack of clustering would not be a result of 'fuzzy' variables. The TOLD-P:2 does have some minor weaknesses as indicated by Word Articulation's low scores for internal consistency ($r = .67–.93$) and stability ($r = .74$), and Grammatical Understanding's low stability ($r = .74$).

The performance IQ variables, Block Design and Picture Competition, were derived from the Wechsler Preschool and Primary Scales of Intelligence—Revised performance scale (WPPSI-R) (Wechsler 1989). The WPPSI-R was developed to test verbal and performance cognitive ability in children. It has decent validity as indicated by high concurrent validity and predictive validity. There were no reported tests of validity for individual subtests. Wechsler reported the stability of each subtest, and both Block Design ($r = .80$) and Picture Completion ($r = .82$) have acceptable stability. Tomblin (2010) reported that the abbreviated scale was a reliable estimate of performance IQ ($r = .73$), therefore any lack of clustering would not be a result of psychometric problems with the variables. Although our choices were limited in terms of possible variables due to the nature of the sample, the EpiSLI language and cognitive variables allow for at least one measure for several language domains (i.e., lexical, grammatical, syntax, pragmatic, phonological, morphosyntax) that would allow for the testing of several different subtypes. Table 2 depicts the relationship between the variables and the language domains.

We examined the distribution of the variables to check for normal distribution using the skew, kurtosis and Shapiro–Wilk (Shapiro and Wilk 1965). The Shapiro–Wilk tests are significant, indicating that the samples deviated from a normal distribution; however, the skew and kurtosis analyses indicated only minor deviations from the normal distribution. Additionally, the ranges included positive and negative values, which indicates that the ranges were not truncated. Table 3 contains ranges, skew, kurtosis and Shapiro–Wilk test results for all variables. Table 4 contains Pearson correlations for the DLD sample with *p*-values presented in the upper triangle. Although some variables are significantly correlated, all correlations are $< 0.50$.

## Analysis

The language and cognitive variables were normalized to ensure that the variables had equal variance and were continuous using Statistical Package for the Social Sciences (SPSS; v. 22.0). The analysis processes are summarized in figure 1 and below. A detailed explanation of the analyses is available in appendix A in the supplemental data online and the R code for these analyses is available in appendix B online. Cluster analyses were conducted in R (2010) using the packages 'cluster' (Maechler *et al.* 2014) and 'matrixStats' (Bengtsson 2010).

We used two clustering techniques: Ward's method and *K*-means. When comparing cluster results from *K*-means and Ward's method, it is not expected that completely identical cluster

results will be obtained, because *K*-means and Ward's methods are based on different algorithms. However, it is expected that results from these two methods will contain highly similar features. Each clustering method was a multi-step process (see figure 1 and appendix A), which involved applying appropriate data transformations, clustering the data, calculating the Bayesian information criteria (BIC; Schwarz 1978) and graphing the results. For the *K*-means analysis, these steps were also completed in a permutated data set. To determine support for a model, we extracted the number of optimal clusters and two measures of degree of randomness. The BIC was used to determine the optimal number of clusters. Degree of randomness was measured using: (1) the spatial distance between clusters; and (2) the distance between BIC curves for real and permutation runs. We used principal components analysis to create the two-dimensional scatter plots for visual inspection of results.

A subtypes model would be validated if the cluster solution has fewer than eight clusters, is spatially distinct, has no overlap with the BIC curve for permutation runs and indicates statistical difference between clusters. We set the upper limit of clusters to eight to prevent subtypes being based on a single test. In our review of the literature, all subtype models had more than one indicator (i.e., behaviour) for classification; therefore, we did not want to yield subtypes that could only be described by one indicator, such as a subtype best described as 'low receptive vocabulary' because this cluster was significantly below all other clusters for Picture Vocabulary but not significantly below for any other measures. An individual differences model would be supported if the cluster solutions are: more than eight clusters or equal to one cluster, lack spatial distinction and the BIC curves overlap. These criteria were selected because they can be used to determine if the cluster solutions occurred randomly; random differences between individuals has been proposed as a feature of an individual differences model in cognitive psychology (Tomblin and Nippold 2014). These criteria were also selected because they map onto the two aspects on which the three models differ: number of clusters and degree of randomness. To support the continuum/spectrum model, the cluster solutions will need to contain more than one cluster, lack a spatial distinction and not overlap with the BIC curve for permutation runs. These criteria were selected because more than one non-distinct cluster would imply a latent structure within the data that is not random; however, the lack of spatial distinction would indicate that the clusters do not substantially differ from one another.

## Results

### Cluster analyses

The BIC curves for *K*-means and Ward's methods are shown in figures 2 and 3 respectively. The minimum BIC corresponded to 18 clusters for both the DLD data set and related permutated data, which was too many clusters to attempt to determine reasonably which features were responsible for the clusters (e.g., clustering of specific language and/or cognitive features). For Ward's method, the cluster size corresponding to the minimum BIC value was 51, which is highlighted in figure 2 by a black box on the dendrogram. The large number of clusters in the DLD data set provides support for both individual differences and continuum/spectrum models, but not for the subtypes model.

Visual analysis of the data yields no spatially distinct clusters (figure 4). The variables were transformed into principal components to reduce the dimensionality of the data, permitting the data to be plotted. The data do not separate into visual distinct clusters in either clustering method. The high number and lack of visual distinct clusters supports the individual differences and continuum/spectrum models, but not the subtypes model.

### Randomization analyses

To test the likelihood that the results were random, bootstrapping and permutations methods were used for *K*-means clustering. This process creates random sets of cases with random values pulled from the database. These sets reflect what would happen if the data were truly random and (most likely) not clustered. If the DLD data set did not contain any clusters, this test would demonstrate how likely we were to find clusters by chance. Mean BIC values, with 95% confidence intervals, for each cluster tested were calculated and compared with clustering results of the original data. As mentioned above, the permutated data had the same minimum cluster size (18) as the DLD cases. However, the BIC values were much larger than the values for the DLD clustering results. The confidence interval envelope (small dashed lines figures 2 and 3) did not overlap with the BIC curve for DLD cases, which means that the results obtained were significantly differ from chance findings. The clusters were unlikely to be due to chance, as represented by the large gap between the associated BIC curves for the *K*-means clustering, as well as lack of overlap with the confidence interval with permutated data (figure 3). The cases with DLD have an underlying structure that is the cause of some overall commonality between cases. The significant difference from chance findings supports the continuum/spectrum model. These findings do not support the individual differences model, because this model would have higher than expected levels of variability and thus be more likely to yield random clustering (i.e., clusters due to chance). The results are summarized in table 5.

### Performance IQ

A secondary purpose of this project was to determine if performance IQ separated children with DLD. In our analyses, no two-cluster solution emerged, so we were unable to complete any analyses to determine if performance IQ separated clusters.

## Discussion

The purpose of this study was to test whether different models of DLD differentially explained the variability in the EpiSLI kindergarten database (Tomblin 2010) using cluster analyses. We tested three models identified on the basis of a priori theoretical views on the nature of DLD: subtypes, individual differences and continuum/spectrum. We attempted to test whether children with DLD formed two groups aggregated by performance IQ ability. The work differs from past research in the following ways: (1) it includes performance IQ measures in the cluster analysis; (2) it does not use predefined groupings; and (3) it compares three competing theoretical models using the same statistical approach. The results support and extend previous cluster analyses (Conti-Ramsden and Botting 1999, Tomblin and Zhang 1999). The results actively reject a subtype model for DLD in the EpiSLI database based on the measures therein. The analyses partially support the individual

differences model, and fully support the continuum/spectrum model. Additionally, there was no indication that performance IQ created distinct groups. We will discuss how these results relate to each of the proposed models and to past research.

The results herein support past research indicating that there are no meaningful subtypes of DLD (Dollaghan 2004, 2011). In our analyses, there were too many cluster solutions and these cluster solutions could not be plausibly mapped onto any a priori subtype models of DLD. Moreover, the optimal number of clusters was $>10$ for both Ward's and $K$-means algorithms. Practically speaking, this large number of clusters indicates that individual clusters did not arise from interpretable feature combinations.

If, by chance, we misidentified the optimal number of clusters based on the BIC, then we should have seen groupings of clusters during visual analysis of the scatterplots, which was not the case. Additionally, if a smaller number of clusters (e.g., four) could have accounted for the all clusters identified (e.g., 51), then this hierarchical structure would have been clearly observable on the dendrogram derived from the Ward's clustering results. Based on the optimal clusters identified and visual analysis of plots, we can confidently reject the subtype model for this database. As outlined in the introduction, past research on subtypes of DLD has had mixed results. The results combined with the results from studies by Dollaghan (2011) and Tomblin and Zhang (1999) strongly suggest, that the subtype model is unlikely to explain variation in clinical presentation of DLD.

We found partial support for an individual differences model. The large number of clusters coupled with a lack of distinction between clusters provide partial support for an individual differences model. On the other hand, the lack of overlap between the two BIC curves indicates that the variability in the EpiSLI sample was *not* predominately random—a key predictive feature of individual differences models that was not evident in our analyses of the EpiSLI database. That is, the high number of 'clusters' supports an individual differences model, but the analyses rejecting randomness in the database do not support this model.

Given the mixed evidence for the individual differences model, it would be premature to discard this model of DLD completely. Commentary in the CATALISE study highlights the need to continue to explore individual differences models: 'there is no clear cut-off that distinguishes between language impairment [ … ] from the lower end of normal variation of language ability' (Bishop *et al.* 2016: 11) and 'many instruments used to assess child language are insensitive to impairments that affect day-to-day language functioning' (12). Stated directly, there is not a sharp group distinction between a relatively low end of a 'typical' sample of language ability and DLD, so it is plausible to argue that DLD, such as typical performance, is relatively equally spaced on a continuum of performance (i.e., with no discernible clusters; Leonard 2014). A direct test of this could be accomplished by replicating the EpiSLI population-based sampling across different age points that accurately captures the range of language development and once again test for clusters and randomness in the sample. Based on our analysis of the EpiSLI database, an individual differences DLD solution remains an open question. We can further investigate the individual differences model by examining whether the variability in presentation is due to random noise by repeating these analyses in other databases by including a wider range of language skills and

instruments that are sensitive to day-to-day functioning, and by modifying the analysis to include typical language peers on all these measures.

We found support for the continuum/spectrum model in the EpiSLI database, which also confirmed our hypothesis. All three criteria a priori outlined in the Methods section were met: (1) a large number of clusters that (2) lacked spatial distinction and (3) were not random. Although *K*-means and Ward's results did not yield precisely the same clusters, both statistical methodologies met the criteria for interpretation as evidence for a spectrum disorder: poor spatial distinction and evidence of non-random clustering (i.e., clusters were not due to chance alone). The results indicated an underlying structure within the data and demonstrated this structure was not due to chance, as demonstrated by the *K*-means clustering and bootstrapping methods (e.g. figure 4).

Additionally, the Ward's methods' results indicated that the structure was unlikely to be hierarchical, especially given the similarity between the *K*-means and Ward's methods results. This underlying, non-random structure is evidence of a spectrum disorder of language abilities. The results support the perspective that children with DLD do not vary by specific symptomology (Reilly *et al.* 2014a, 2014b), but form a cohesive diagnostic group. Thus, the data are fully aligned with the continuum/spectrum model; there was a high number of clusters (which could not be mapped onto subtypes) *and* the data were *not* randomly distributed. This then implies cluster around severity of key DLD traits rather than clinically meaningful subtypes.

Moreover, a continuum/spectrum perspective recognizes that the focus of assessment should be differential diagnosis of severity *and* individual patterns of strengths and weaknesses (as is also consistent with individual differences) and treatment plans should focus on severity *in addition to* individual traits. This perspective does align with the clinical implications of an individual differences model while also recognizing that different severity levels can motivate different treatment paths (and research paradigms).

## Relationship to previous studies of subtypes in DLD

The results can plausibly account for the inconsistent past research on subtypes as a continuum/spectrum disorder can be inadvertently 'forced' into subtypes when applying a univariate approach to define each subtype. Most successful subtype research focused on creating subtypes based on a single characteristic per group, such as grammatical SLI (Bishop *et al.* 2000a). This is because within a spectrum condition it is not unlikely at all to find children who are low on one skill (e.g., tense marking) and high on another (e.g., social appropriateness) to create 'confirmatory' groups. Non-random or even random selection from the larger DLD population with a continuum/spectrum structure could result in positive evidence for subtypes, because there can be enough underlying structure to form related groupings. At that point sampling error can mistakenly support that there are two subgroups even though these do not exist in the actual DLD population. If this were the case, then these confirmatory 'subgroup' analyses would not be systematically replicable. That is, when researchers attempt to replicate those results or to define subtypes based on more than one dimension (e.g., impaired vocabulary and grammatical subtype) the research becomes inconsistent and unstable.

As an example, Conti-Ramsden *et al.* (1997) successfully found subtypes using cluster analysis. But, when this same research team re-examined the children at age 8 years, 45% of the children had changed clusters (Conti-Ramsden and Botting 1999). This finding indicates instability of subtypes, which is exactly what we would predict if the DLD population was a continuum/spectrum that was forced into subtypes within a univariate framework. DLD is not the first condition to move from subtypes to a continuum/spectrum approach. Although novel in the area of DLD, this is following in the well-worn path of arguments for (and against) reconceptualizing ASD (Lord and Risi 1998, Wing 1997) and conditions such as ADHD and LD (Mayes *et al.* 2000) as spectrum conditions rather than subtypes or individual differences. For example, autism sharply shifted from subtypes of PDD in the DSM-IV to ASD in DSM-V (Frazier *et al.* 2012). Although nescient, the results herein suggest that DLD be systematically tested within a spectrum framework in future population-based samples and that the field engage in a thoughtful debate as to whether a spectrum approach to DLD is warranted.

## Limitations

There are two main limitations to this project. These arise from the nature of the database and the nature of cluster analysis. The EpiSLI database is one of a few comprehensive epidemiological samples representative of the United States, which provides a unique opportunity to explore subtypes of DLD. The limitation to using the EpiSLI is that, due to constraints surrounding the original project, only a limited number of language and cognitive variables were assessed at *one developmental age* (kindergarten). One example of how the included variables limit our interpretation is that the Culatta tasks (e.g., narrative retell and narrative comprehension) are not similar to norm-referenced pragmatic language tasks. Therefore, our analysis was limited in its ability to test for the presence or absence of a pragmatic language disorder subtype because of the nature of the variables included. This is a limitation that can be easily addressed in future research by including more language measures overall and more tasks within a domain.

Additionally, cluster analysis approaches themselves have limitations. Cluster analysis is inherently probabilistic and thus subject to variance in specific numbers of clusters identified in any particular run. As our results demonstrate, different clustering methods will predictably yield different cluster results, which is why it is crucial not to rely on a single-cluster analysis result to draw firm conclusions. We addressed this known limitation by using both hierarchical and agglomerative methods, as well as using principal components analysis. One argument against clustering methods is that cluster analysis works best in samples where any potential overlap between clusters is either absent or minimal (Beauchaine 2003, 2007). However, it is appropriate to apply statistical cluster analysis methods that could identify if there were stable non-overlapping groupings as well as random and unstable groups, which cluster analysis did allow us to test. Furthermore, we wanted to impose structure on the sample and then determine if that structure reliable fit with one of the three identified models.

### Future directions

We encourage replication of these results using alternative clustering methods or other statistical analyses, such as factor analysis. However, future studies should use a larger, population-based database that contains newer and more comprehensive measures and that spans more than one developmental age (Raghavan *et al.* 2018). New databases will need to include a more diverse sampling of children with DLD, especially across cultures and languages. A second avenue of research is investigating the core features of DLD. Others have argued that updated population studies are needed in DLD and other speech and/or language disabilities (Raghavan *et al.* 2018).

Previous work has highlighted deficits in grammar (Rice *et al.* 1998, 2004, Hadley and Holt 2006), syntax (Nippold *et al.* 2009) and working memory (Ellis Weismer *et al.* 1999, Montgomery, Magimairaj, and Finney 2010, Henry and Botting 2017). The CATALISE study provides a list of several principal areas for intervention, including phonology, syntax, pragmatics, discourse, word finding and verbal learning (Bishop *et al.* 2017), which should be considered as possible core features. Based on recent research on the dimensionality of language in kindergarten (Tomblin and Zhang 2006, LARRC 2015), it is possible that core features may not relate to domains of language but rather with proficiency and acquisition. Additionally, it is possible that the core features might change over time as language develops, which would account for research suggesting that the dimensionality of language changes over time (Lonigan and Milburn 2017). Future researchers investigating the validity of the individual differences model will want to include children with typical language in an appropriate case to control ratio.

### Clinical implications

The clinical recommendations for a continuum/spectrum disorder are limited currently due to the brevity of research; however, combining our work with research in other disorders (e.g., autism) and the theoretical underpinnings of DLD, we can say that a continuum/spectrum approach has implications for the focus of assessment and developmental of treatment plans. Should the continuum/spectrum approach be upheld after more research the focus of assessment should be differential diagnosis and treatment plans should focus on individual traits and severity. We can speculate that many clinicians already engage in this approach, but implementation studies could determine to what extent these procedures are used in current practice and whether combining severity and trait analyses improves outcomes. A continuum/spectrum model would support the practice of developing individualized treatment plans and would raise a demand for development of new assessment procedures focused on core elements and functional capacity.

## Conclusions

The evidence is stronger for the continuum/spectrum model because this model best fits the EpiSLI database (one of the sole population-based samples of DLD in the United States) and helps refine the interpretation of inconsistent findings from previous research. As with previous conceptualizations of ASD and ADHD, there do not appear to be recognizable (and thus diagnosable) subtypes in DLD. In contrast, we cannot completely reject the individual

differences model because we did find some positive evidence. Given the support for continuum/spectrum model in this study, we encourage researchers to focus on identifying the core symptomology features of DLD. Past research has already highlighted grammar and syntax as probable core symptoms. It is important clinically and diagnostically to continue to look at the individual differences and continuum/spectrum models, as these models will help to determine best clinical practices and provide direction for research examining the genetics and neurological aspects of DLD.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

Acknowledgement is given to the original grant (#N01-DC-1–2107) and supplement (#3 P50 DC002746–08S1) from the National Institute on Deafness and Other Communication Disorders, a division of the National Institutes of Health. Thanks also to those who provided constructive feedback on the manuscript. This study was supported by a Preparation of Leadership Personnel grant (H325D080075; PI: Schuele Co-PI: Camarata) from the US Department of Education and by the Scottish Rite Foundation of Nashville.

## References

AMERICAN PSYCHIATRIC ASSOCIATION, 1980, Diagnostic and Statistical Manual of Mental Disorders (3rd ed.) (Washington D.C.: American Psychiatric Association).

AMERICAN PSYCHIATRIC ASSOCIATION, 1994, Diagnostic and Statistical Manual of Mental Disorders (4th ed.) (Washington D.C.: American Psychiatric Association).

AMERICAN PSYCHIATRIC ASSOCIATION, 2013, Diagnositic and Statistical Manual of Mental Disorders (5th ed.) (Washington D.C: American Psychiatric Association).

ARAM DM and NATION JE, 1975, Patterns of language behavior in children with developmental language disorders. Journal of Speech, Language, and Hearing Research, 18(2), 229–241.

BEAUCHAINE TP, 2003, Taxometrics and developmental psychopathology. Development and Psychopathology, 15(3), 501–527. 10.1017/S0954579403000270 [PubMed: 14582930]

BEAUCHAINE TP, 2007, A brief taxometrics primer. Journal of Clinical Child and Adolescent Psychology, 36, 37–41. 10.1080/15374410701662840.A

BENGTSSON H, 2010, Matrixstats: Methods that apply to rows and columns of a matrix

BISHOP DVM, 1997, Uncommon Understanding: Development and Disorders in Language Comprehension in Children (Cambridge, UK: Psychology Press).

BISHOP DVM, 2004, Specific language impairment: diagnostic dilemmas. In Verhoeven LT and Balkom H (eds), Classification of Developmental Language Disorders: Theoretical Issues and Clinical Implications (Mahwah, NJ: Lawrence Erlbaum Association), pp. 309–326.

BISHOP DVM, 2017, Why is it so hard to reach agreement on terminology? The case of developmental language disorder (DLD). International Journal of Language and Communication Disorders, 52(6), 671–680. 10.1111/1460-6984.12335 [PubMed: 28714100]

BISHOP DVM, BRIGHT P, JAMES C, BISHOP SJ and VAN DER LELY HKJ, 2000a, Grammatical SLI: a distinct subtype of developmental language impairment. Applied Psycholinguistics, 21, 159–181.

BISHOP DVM, CHAN J, ADAMS CM, HRTLEY J and WEIR F, 2000b, Converstational responses in specific language impairment: evidence of disproportionate pragmatic difficulties in a subset of children. Development and Psychopathology, 12, 177–199. [PubMed: 10847623]

BISHOP DVM, SNOWLING MJ, THOMPSON PA, GREENHALGH T and CATALISE CONSORTIUM, 2016, CATALISE: a multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. Plos One, 11(7). 10.7287/PEERJ.PREPRINTS.1986

BISHOP DVM, SNOWLING MJ, THOMPSON PA, GREENHALGH T and CATALISE-2 CONSORTIUM, 2017, Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: terminology. Journal of Child Psychology and Psychiatry, 58(10), 1068–1080. 10.1111/jcpp.12721 [PubMed: 28369935]

BISHOP DVM and ROSENBLOOM L, 1987, Childhood language disorders: classification and overview. Language Development and Disorders, 101, 16–41.

BOTTING N and CONTI-RAMSDEN G, 2004, Characteristics of children with specific language impairment. In Verhoeven LT and Balkom H (eds), Classification of Developmental Language Disorders: Theoretical Issues and Clinical Implications (Mahwah, NJ: Lawrence Erlbaum Association), pp. 23–38.

CAMARATA S and SWISHER L, 1990, A note on intelligence assessment within studies of specific language impairment. Journal of Speech, Language, and Hearing Research, 33(1), 205–207.

CONTI-RAMSDEN G and BOTTING N, 1999, Classification of children with specific language impairment: longitudinal considerations. Journal of Speech, Language and Hearing Research, 42, 1195–1204.

CONTI-RAMSDEN G, CRUTCHLEY A and BOTTING N, 1997, The extent to which psychometric differentiate subgroups of children with SLI. Journal of Speech, Language and Hearing Research, 40, 765–777.

CULATTA B, PAGE JL and ELLIS J, 1983, Story telling as a communicative performance screening tool. Language, Speech, and Hearing Services in Schools, 14, 66–74.

DOLLAGHAN CA, 2004, Taxometric analyses of specific language impairment in 3- and 4-year-old children. Journal of Speech, Language and Hearing Research, 47, 464–475.

DOLLAGHAN CA, 2011, Taxometric analyses of specific language impairment in 6-year-old children. Journal of Speech, Language and Hearing Research, 54, 1361–1371.

ELLIS WEISMER S, EVANS JL and HESKETH LJ, 1999, An examination of verbal working memory capacity in children with specific language impairment. Journal of Speech, Language and Hearing Research, 42, 1249–1260.

FEAGANS L, PORTER F, CHILD G and APPLEBAUM MI, 1986, Validation of language subtypes in learning disabled children. Journal of Educational Psychology, 78(5), 358–364. 10.1037/0022-0663.78.5.358

FRAZIER TW, YOUNGSTROM EA, SPEER L, EMBACHER R, LAW P, CONSTANTINO J, FINDLING RL HARDAN AY and ENG C, 2012, Validation of proposed DSM-5 criteria for autism spectrum disorder. Journal of the American Academy of Child and Adolescent Psychiatry, 51(1), 28–40. 10.1016/j.jaac.2011.09.021 [PubMed: 22176937]

HADLEY PA and HOLT JK, 2006, Individual differences in the onset of tense marking: a growth-curve analysis. Journal of Speech, Language and Hearing Research, 49, 984–1000.

HENRY LA and BOTTING N, 2017, Working memory and developmental language impairments. Child Language Teaching and Therapy, 33(1), 19–32. 10.1177/0265659016655378

KANNER L, 1968, Autistic disturbances of affective contct. Acta Paedopsychiatrica, 35(4), 100. [PubMed: 4880460]

LARRC, 2015, The dimensionality of language ability in young children. Child Development, 86(6), 1948–1965. 10.1111/cdev.12450 [PubMed: 26509742]

LEONARD LB, 1989, Language learnability and specific language impairment in children. Applied Psycholinguistics, 10, 179–202.

LEONARD LB, 2009, Is expressive language disorder and accurate diagnostic category. American Journal of Speech–Language Pathology, 18, 115–123. [PubMed: 19029534]

LEONARD LB, 2010, Language combinations, subtypes, and severity in the study of bilingual children with specific language impairment. Applied Psycholinguistics, 31, 310–315. [PubMed: 23576828]

LEONARD LB, 2014, Children with Specific Language Impairment, 2nd edn (Cambridge, MA: MIT Press).

LONIGAN CJ and MILBURN TF, 2017, Identifying the dimensionality of oral language skills of children with typical development in preschool through fifth grade. Journal of Speech, Language and Hearing Research, 60(8), 2185–2198. 10.1044/2017_JSLHR-L-15-0402

LORD C and RISI S, 1998, Frameworks and methods in diagnosing autism spectrum disorders. Mental Retardation and Developmental Disabilities Research Reviews, 4(2), 90–96. 10.1002/(SICI)1098-2779(1998)4:2<90::AID-MRDD5>3.0.CO;2-0

MACKIE C and DOCKRELL JE, 2004, Deficits in children with SLI. Hearing Research, 47, 1469–1484.

MAECHLER M, ROUSSEEUW P, STRUYF A and HUBERT M, 2014, Cluster: Cluster analysis basics and extensions: R package version 1.15.3

MAYES SD, CALHOUN SL and CROWELL EW, 2000, Learning disabilities and ADHD: overlapping spectrum disorders. Journal of Learning Disabilities, 33(5), 417–424. 10.1177/002221940003300502 [PubMed: 15495544]

MONTGOMERY JW, MAGIMAIRAJ BM and FINNEY MC, 2010, Working memory and specific language impairment: an update on the relation and perspectives on assessment and treatment, American Journal of Speech Language Pathology, 19, 78–94. 10.1044/1058-0360(2009/09-0028) [PubMed: 19948760]

NEWCOMER P and HAMMILL D, 1988, Test of Language Development—Primary 2nd Edition (Austin, TX: PRO-ED).

NIPPOLD MA, MANSFILED TC, BILLOW JL and TOMBLIN JB, 2009, Syntactic development in adolescents with a history of language impairments: a follow-up investigation. American Journal of Speech–Language Pathology, 18, 241–251 (available at: http://ajslp.asha.org/cgi/content/abstract/18/3/241) (accessed on 16 June 2012). [PubMed: 19106210]

PRIOR M, EISENMAJER R, LEEKAM S, WING L, GOULD J, OG B and DOWE D, 1998, Are there subgroups within the autistic spectrum? A cluster analysis of a group of children with autistic spectrum disorders. Journal of Child Psychology and Psychiatry, 39(6), 893–902 (available at: http://www.ncbi.nlm.nih.gov/pubmed/9758197). [PubMed: 9758197]

RAGHAVAN R, CAMARATA S, WHITE K, BARBARESI W, PARISH S and KRAHN G, 2018, Population health in pediatric speech and language disorders: Data sources and a research agenda for the field. Journal of Speech, Language, and Hearing Research, 61(5), 1279–1291.

RAPIN I and ALLEN DA, 1987, Developmental dysphasia and autism in preschool children: characteristics and subtypes, in Proceedings of the First International Symposium on Specific Speech and Language Disorders in Children (London: AFA-SIC), pp. 20–35.

REILLY S, BISHOP DVM and TOMBLIN JB, 2014a, Terminological debate over language impairment in children: forward movement and sticking points. International Journal of Language and Communication Disorders, 49(4), 452–462. 10.1111/1460-6984.12111 [PubMed: 25142092]

REILLY S, TOMBLIN JB, LW J, MCKEAN C, MENSAH FK, MORGN A, GOLDFELD S, NICHOLSON JM and WAKE M, 2014b, Specific language impairment: a convenient label for whom? International Journal of Language and Communication Disorders, 49(4), 416–451. 10.1111/1460-6984.12102 [PubMed: 25142091]

RICE ML, TOMBLIN JB, RICHMN WA and MARQUIS J, 2004, Grammatical tense deficits in children with SLI and nonspecific language impairment: relationships with nonverbal IQ over time. Journal of Speech, Language and Hearing Research, 47, 816–834.

RICE ML, WEXLER K and HERSHBERGER S, 1998, Tense over time: the longitudinal course of tense acquisition in children with specific language impairment. Journal of Speech, Language and Hearing Research, 41, 1412–1431.

SHAPIRO S and WILK M, 1965, An analysis of variance test for normality (complete samples). Biometrika, 52(3/4), 591–611.

SCHWARZ G, 1978, Estimating the dimension of a model. The Annuals of Statistics, 6, 461–464.

STARK R and TALLAL P, 1981, Selection of children with specific language deficits. Journal of Speech, Language, and Hearing Research, 46(2), 114–122.

TOMBLIN JB, 2010, The EpiSLI database: a publicly available database on speech and language. Language, Speech, and Hearing Services in Schools, 41(1), 108–117. 10.1044/0161-1461(2009/08-0057)

TOMBLIN JB, RECORDS NL, BUCKWALTER P, SMITH E and O'BRIEN M, 1997, Prevalence of specific language impairment in kindergarten children. Journal of Speech, Language and Hearing Research, 40, 1245–1260.

TOMBLIN JB and NIPPOLD MA, 2014, Understanding Individual Differences in Language Development across the School Years (New York, NY: Psychology Press).

TOMBLIN JB, RECORDS NL and ZHANG X, 1996, A system for the diagnosis of specific language impairment in kindergarten children. Journal of Speech and Hearing Research, 39(6), 1284–1294 (available at: http://www.ncbi.nlm.nih.gov/pubmed/8959613). [PubMed: 8959613]

TOMBLIN JB and ZHANG X, 1999, Language patterns and etiology in children with specific language impairment. In Tager-Flusberg H (ed.), Neurodevelopmental Disorders: Contributions to a New Framework from the Cognitive Neurosciences (Cambridge, MA: MIT Press), pp. 361–382.

TOMBLIN JB and ZHANG X, 2006, The dimensionality of language ability in school-age children. Journal of Speech, Language and Hearing Research, 49, 1193–1208 (available at: http://jslhr.highwire.org/cgi/content/abstract/49/6/1193) (accessed on 16 June 2012).

WECHSLER D, 1989, Wechsler Preschool and Primary Scale of Intelligence—Revised (San Antonio, TX: Psychological Corporation).

WIGGINS LD et al., 2012, Support for a dimensional view of autism spectrum disorders in toddlers. Journal of Autism and Developmental Disorders, 42(2), 191–200. 10.1007/s10803-011-1230-0 [PubMed: 21448751]

WING L, 1997, The autistic spectrum. Lancet, 350(9093), 1761–1766. 10.1016/S0140-6736(97)09218-0 [PubMed: 9413479]

ZAIHRA T, WALSH CJ, AHMED S, FUGERE C, HAMID Q, OLIVENSTEIN R, MARTIN JG and BENEDETTI A, 2016, Phenotyping of difficult asthma using longitudinal physiological and biomarker measurements reveals significant differences in stability between clusters. BMC Pulmonary Medicine, 16(1), 74 10.1186/s12890-016-0232-2 [PubMed: 27165150]

## What this paper adds

**What is already known on the subject**

The DLD population is highly heterogeneous. This heterogeneity attenuates our ability to determine whether the DLD population is (1) made up of smaller homogeneous 'subgroups'; or (2) represents the lower end of normal language ability as a form of individual differences; or (3) like autism spectrum disorder (ASD), is a spectrum disorder.

**What this paper adds to existing knowledge**

This paper compares statistically three models currently debated in the DLD literature to examine which best fits the underlying population structure of a US-based epidemiological sample of children with DLD.

**What are the potential or actual clinical implications of this work?**

There are at least two clinical implications. First, some have suggested that DLD includes specific distinct subtypes, but the results herein suggest that clinically interpretable subtypes are unlikely (based on the measures used in the current analysis). Second, we propose that a reconceptualization of DLD as a 'spectrum' condition for the purposes of diagnosis, intervention and phenotyping should be explored in more detail in clinical samples and future studies.

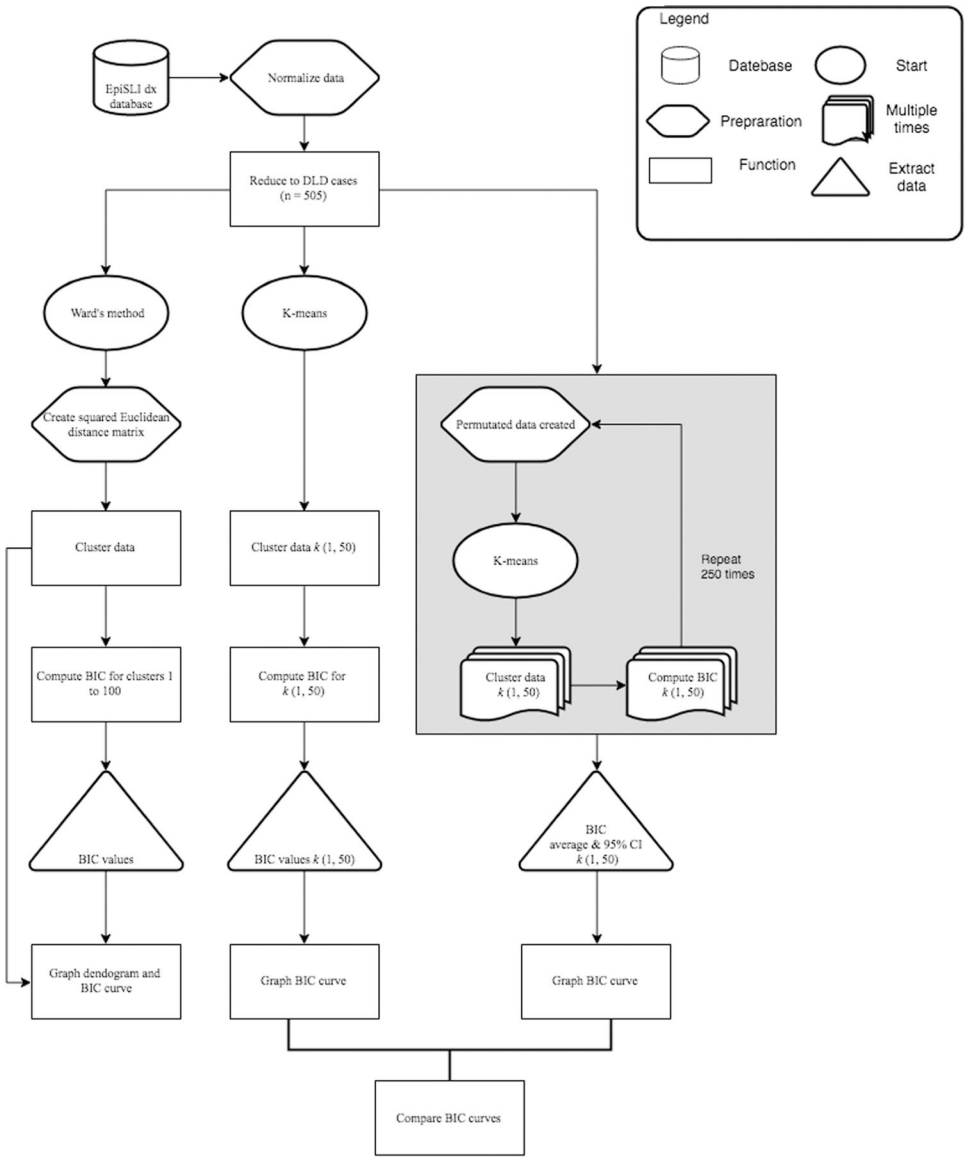


**Figure 1.**
Flow chart summary of the analysis.

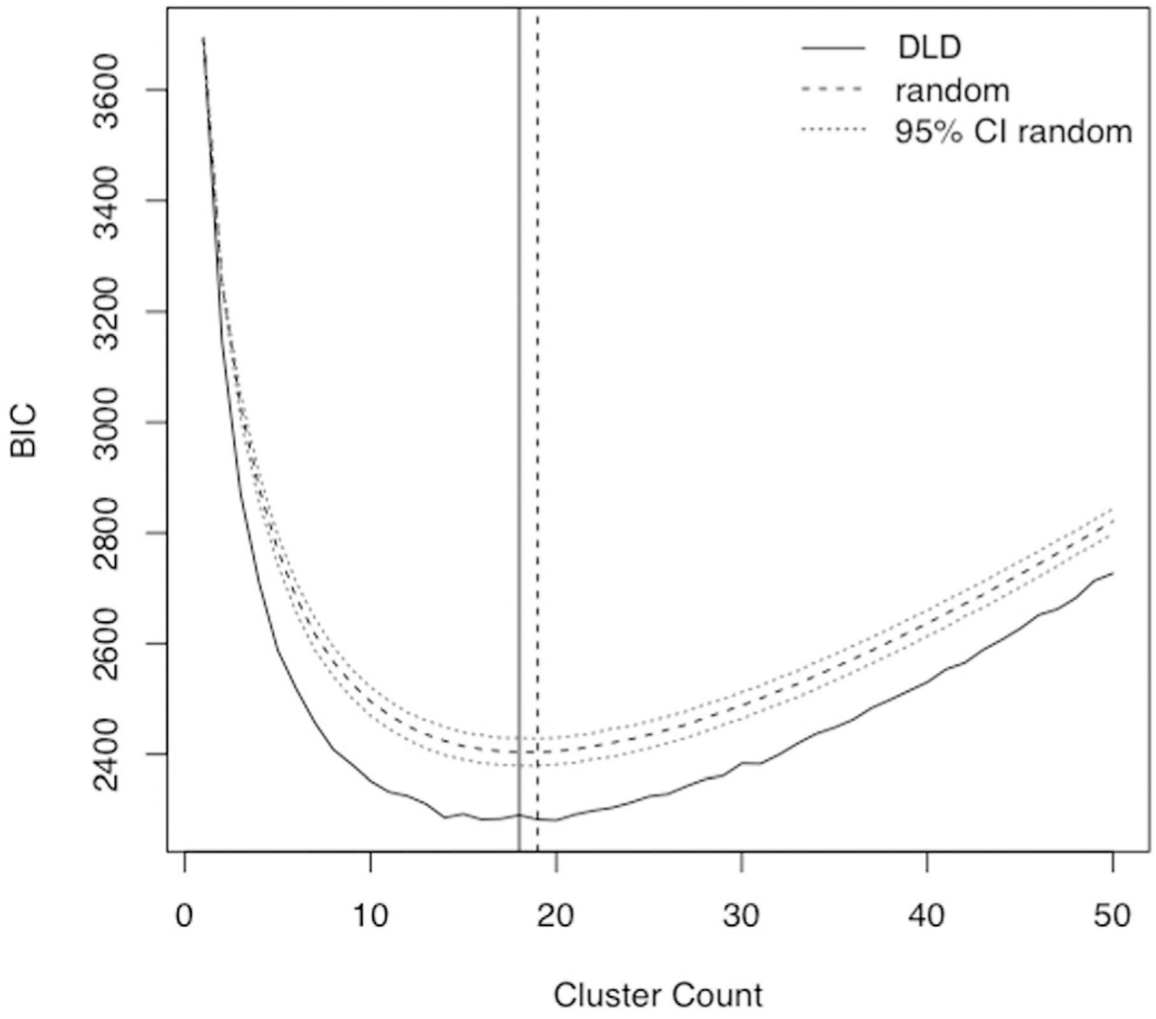


**Figure 2.**

Bayesian information criteria (BIC) curves for *K*-means clustering for the data set. BIC curves are for the original data, permutated data and 95% confidence interval envelope for the permutated curve.

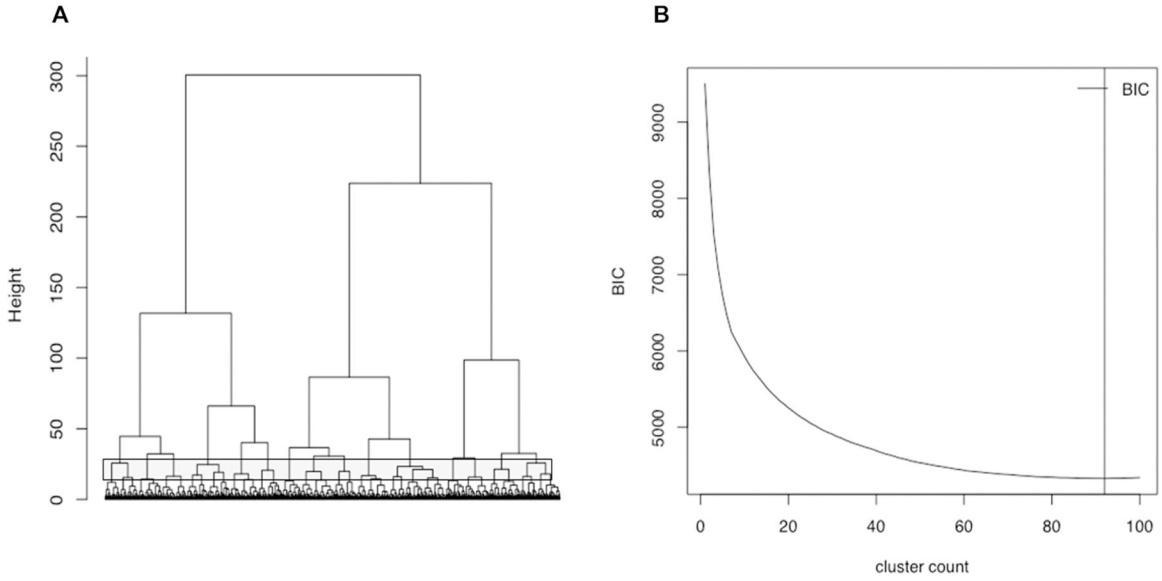


**Figure 3.**

Ward's hierarchical clustering results for the data set. (A) The dendrogram for the hierarchical, where the *y*-axis is height, which is the criterion value for a merge and represents the total distance accounted for by that merge. The black box on the dendrogram represents the identified minimum Bayesian information criteria (BIC) value. (B) Associated BIC curve for clusters 1–100 for the data set. The vertical line represents the identified minimum BIC value.

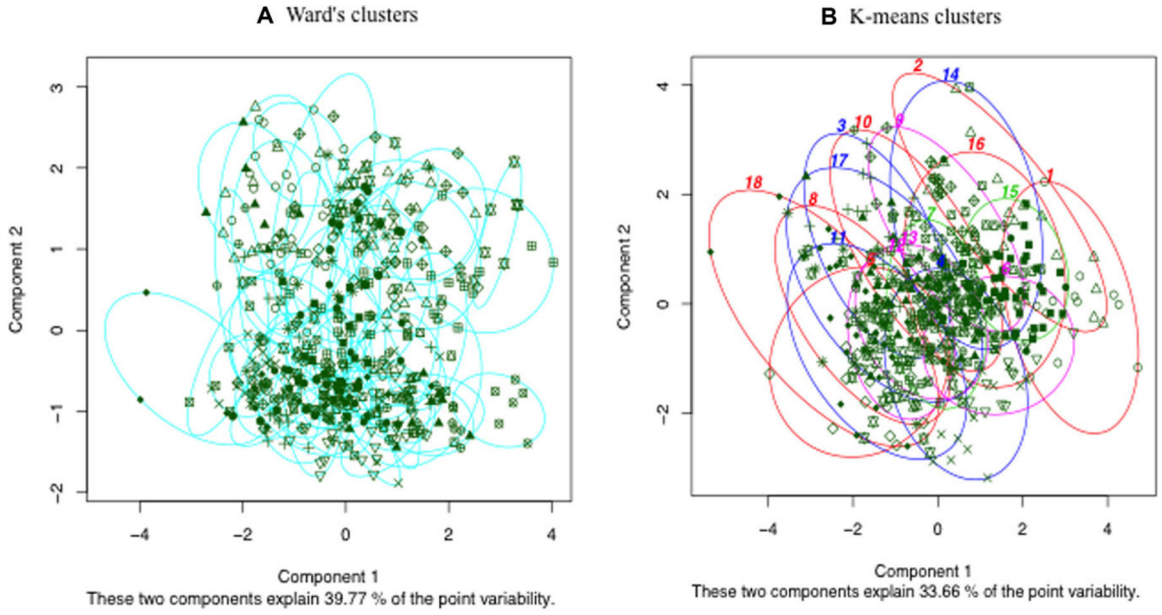


**Figure 4.**

Ward's method cluster scatter plot (A) and *K*-means cluster scatterplot (B) for the developmental language disorder (DLD) data set. Circles represent clusters. [Colour figure can be viewed at wileyonlinelibrary.com]

**Table 1.**

How the models relate to the degree of randomness and the number of clusters

| | Number of clusters | |
|---|---|---|
| **Degree of randomness** | **Low** | **High** |
| High | No interpretable model | Individual differences |
| Low | Clinically meaningful subtypes | Continuum/spectrum |

**Table 2.**

Relationship between variables and language domains

| | Domains | | | | | |
|---|---|---|---|---|---|---|
| **Variable** | **Lexical** | **Grammatical** | **Syntax** | **Pragmatic** | **Phonological** | **Morphosyntax** |
| Picture Vocabulary | * | | | | | |
| Oral Vocabulary | * | * | * | | | |
| Grammatical Completion | | * | | | | |
| Grammatical Understanding | | * | * | | | * |
| Sentence Imitation | | | | | * | |
| Word Articulation | | | | | * | |
| Narrative recall | | | * | * | | |
| Narrative comprehension | | | | * | | |
| Block Design | | | | | | |
| Picture Completion | * | | | | | |

*Int J Lang Commun Disord*. Author manuscript; available in PMC 2019 August 06.

**Table 3.**

Descriptive information for children with language impairment in the EpiSLI database

| *Demographics* | | | | |
|---|---|---|---|---|
| Age | 72 months (3.8 months) | | | |
| Gender | 56% male | | | |
| Race/ethnicity | 76.2% Caucasian | | | |
| *Language measures* | Mean (standard deviation) [minimum, maximum] | Skew | Kurtosis | Shapiro–Wilk |
| TOLD-P:2—Picture Vocabulary | −0.76 (0.87) [−2.88, 1.69] | 0.12 | −0.32 | 0.991** |
| TOLD-P:2—Oral Vocabulary | −0.85 (0.58) [−1.6, 1.79] | 0.93 | 1 | 0.928** |
| TOLD-P:2—Grammatical Understanding | −0.82 (0.93) [−3.47, 1.89] | −0.18 | −0.27 | 0.989** |
| TOLD-P:2—Grammatical Completion | −0.96 (0.66) [−1.84, 1.46] | 0.64 | −0.19 | 0.943** |
| TOLD-P:2—Sentence Imitation | −0.88 (0.48) [−1.54, 1.05] | 1.34 | 2.16 | 0.889** |
| TOLD-P:2—Word Articulation | −0.29 (1.06) [−2.37, 0.78] | −0.61 | −1.06 | 0.859** |
| Narrative recall | −0.75 (0.74) [−1.74, 1.29] | 0.63 | −0.28 | 0.936** |
| Narrative comprehension | −0.98 (0.99) [−3.72, 1.12] | −0.29 | −0.09 | 0.965** |
| *Cognitive measures* | | | | |
| WPPSI-R—Block Design | –0.52 (0.98) [–2.7, 2.2] | 0.11 | –0.34 | 0.984** |
| WPPSI-R—Picture Completion | –0.50 (1.02) [–3.12, 2.35] | 0.05 | –0.26 | 0.986** |

Notes: TOLD-P:2 = Test of Language Development—Primary 2nd Edition (Newcomer and Hammill 1988); WPPSI-R = Wechsler Preschool and Primary Scales of Intelligence—Revised performance scale (Wechsler 1989).

** $p < 0.001$.

**Table 4.**

Correlations between variation for children with developmental language disorder (DLD) in the EpiSLI database

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. Picture Completion | . | 0.0000 | 0.2274 | 1.0000 | 1.0000 | 1.0000 | 0.0001 | 0.0031 | 0.2841 | 0.0018 |
| 2. Block Design | 0.47 | . | 1.0000 | 1.0000 | 1.0000 | 0.4628 | 0.0000 | 0.0697 | 0.0196 | 0.0174 |
| 3. Narrative Comprehension | 0.12 | 0.06 | . | 0.0015 | 1.0000 | 1.0000 | 1.0000 | 0.0435 | 1.0000 | 0.5209 |
| 4. Narrative Recall | 0.00 | 0.02 | 0.18 | . | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.0435 | 1.0000 |
| 5. Word Articulation | 0.06 | 0.04 | −0.05 | 0.05 | . | 0.0016 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 6. Sentence Imitation | 0.07 | 0.10 | 0.02 | 0.00 | 0.18 | . | 0.0697 | 0.0000 | 0.0742 | 1.0000 |
| 7. Grammatical Understanding | 0.21 | 0.25 | 0.03 | −0.05 | 0.00 | 0.13 | . | 0.3268 | 0.0486 | 0.0361 |
| 8. Grammatical Completion | 0.17 | 0.13 | 0.14 | 0.04 | 0.06 | 0.36 | 0.11 | . | 0.0002 | 0.0466 |
| 9. Oral Vocabulary | 0.11 | 0.15 | 0.07 | 0.14 | 0.01 | 0.13 | 0.14 | 0.20 | . | 1.0000 |
| 10. Picture Vocabulary | 0.18 | 0.15 | 0.10 | −0.06 | 0.04 | 0.04 | 0.15 | 0.14 | 0.07 | . |

Note: Column numbers correspond to row numbers. Values in the upper right triangle are *p*-values.

*Int J Lang Commun Disord*. Author manuscript; available in PMC 2019 August 06.

**Table 5.**

Summary of cluster results and how findings support the different models

| | Finding supports | | |
|---|---|---|---|
| **Finding** | **Subtype** | **Individual differences** | **Continuum/spectrum** |
| Large number of clusters | No | Yes | Yes |
| No overlap between the Bayesian information criteria (BIC) curves (degree of randomness) | Yes | No | Yes |
| No spatial distinction (degree of randomness) | No | Yes | Yes |