Language disorder typologies: Clustering and Principal Component Analysis in the EpiSLI database

By

Hope Sparks Lancaster

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partially fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Hearing and Speech Sciences

May 2015

Nashville, Tennessee

Approved:

Stephen Camarata, Ph.D., Chair

Daniel Ashmead, Ph.D.

David Lubinski, Ph.D.

Matt Shotwell, Ph.D.

ACKNOWLEDGEMENTS

This project would not have been possible without the financial support of US Department of Education Preparation of Leadership Personnel grant (H325D080075; PI: Schuele). Additional acknowledgement is given to 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. I owe much to my dissertation committee, especially Dr.s Camarata and Shotwell, for their guidance on this project.

TABLE OF CONTENTS

Appendix

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DSM-5	Diagnostic and Statistical Manual of Mental Disorders 5th edition
APA	American Psychiatric Association
EpiSLI	Epidemiological study of Specific Language Impairment
NIDCD	National Institute for Deafness and Other Communication Disorders
SLI	Specific language impairment
NLI	Nonspecific language impairment
TOLD-P:2	Test of Language Development Primary 2nd edition
WPPSI-R	Weschler Preschool and Primary Scales of Intelligence Revised
SPSS	Statistical Package for the Social Sciences
BIC	Bayesian Information Criterion
PCA	Principal Components Analysis

CHAPTER 1

INTRODUCTION

Roughly seven percent of children struggle to understand and produce language in the absence of a known primary etiology, such as hearing loss or global intellectual deficit (Tomblin et al., 1997). From a public health perspective, over a lifetime, children with language impairment face many challenges, academically, socially, and economically relative to peers without language impairment (Bishop, 1997; Leonard, 2014; Tomblin, 2008; Tomblin & Nippold, 2014; Watkins & Rice, 1994). There are many possible causes for language impairment of this nature including genetic contributions (Bishop, 1997a), auditory processing deficits (Tallal, 1976), and working memory limitations (Montgomery, 2002; Montgomery, Magimairaj, & Finney, 2010); however none of the proposed theories explain the heterogeneity and individual differences in language impairment (Leonard, 2014; Watkins & Rice, 1994).

Research seeking to understand language impairment, and its correlates, has a long history. In 1892, Gall presented one of the first descriptions of children with language difficulties with no obvious cause (as cited in, Leonard, 2014). Gall reported that the children did not exhibit the same capacity for language as their peers, despite intact hearing and cognition. Numerous case studies followed during the 19th century and described atypical development of language (cf. Leonard, 2014). Researchers use many different terms to describe children with language impairment, such as developmental aphasia, dyspraxia, language disorder, and specific language impairment (cf. Leonard, 2014). Over the past several decades, researchers discovered important characteristics of language impairment and developed more accurate methods for identifying language impairment and other conditions such as global intellectual deficits and autism (e.g., Leonard, 2014). However, it has also become clear that there is substantial heterogeneity in children with language impairment (cf.

Leonard, 2014; Watkins & Rice, 1994). For example, some children with expressive language impairment have relatively normal comprehension of spoken language whereas others display significant weaknesses in both the receptive and the expressive modalities. This heterogeneity has resulted in divergent perspectives on whether there is relatively random variation across traits or whether there are distinct subtypes.

Currently, the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; *DSM-5*; American Psychiatric Association [APA], 2013) defines language impairment as a deficit in a child's ability to understand and produce language in the absence of a primary etiology not attributable to the language deficit (e.g., hearing loss; autism; intellectual disability). The *DSM-5* definition focuses on grammar, vocabulary, and pragmatics; however others have included several additional domains of language and language impairment (Bishop, 1997; Newcomer & Hammill, 1988). The domains are: (a) vocabulary, (b) phonology, (c) syntax, (d) grammar, (e) pragmatics, and (f) morphology. All of the domains have expressive and receptive domains. **Vocabulary** is the number of words that a child can produce and/or comprehend. An example of vocabulary deficit is word-finding problems (Leonard, 2000). **Phonology** is the process of producing and comprehending the speech sounds of a language (e.g., articulation). **Syntax** is the process for combining clauses, phrase and sentences. Children with language impairment show a delay in the development of and errors with complex syntax (e.g., connecting clauses; Nippold, Mansfield, Billow, & Tomblin, 2008). **Grammar** is the set of rules for how words are used in a language. Children with language impairments show a delay in developing appropriate grammar skills and continue to make grammatical errors throughout their life. An example of grammatical errors is omission of tense marking (e.g., past tense *–ed*; Leonard et al, 2003) which is a consistent clinical maker of language impairment (Rice, Wexler, & Hershberger, 1998). **Pragmatics** is the social rules of language (e.g., turn taking). **Morphology** is the structure of words and word formation.

Subtypes of Language Impairment

The literature on language impairment indicates a substantial amount of heterogeneity with regard to the symptomology and outcomes in children with language impairment (Bishop, 1997a; Leonard, 2014). The heterogeneity has led to the question of subtypes of language impairment (e.g., Bishop & Rosenbloom, 1987; Leonard, 2014; Rapin & Allen, 1983; Tomblin et al., 1997). Subtypes could provide clinicians and researchers with valuable information to better direct both research and clinical practice by providing a profile of strengths and weaknesses for a subtype. The Office of Special Education Programs (OSPEP; Data Accountability Center, 2012) recently highlighted the need to improve identification of children with language impairment, and specifically the need to determine if there are replicable subtypes of language impairment. For example, children who show distinct deficits in grammar and syntax, but not vocabulary, might represent a subtype, whereas a different subtype could be characterized by deficits in all three domains. It is certainly possible, at least from a theoretical perspective that children in the different subtypes would potentially benefit from different intervention strategies.

The reported heterogeneity within the population of children with language impairment could be the result of either systematic variation or individual differences (Tomblin & Nipplod, 2014). Individual differences are expected to occur in any large population regardless of whether subtypes are evident. But, against the larger backdrop of individual differences, regularities, or clusters within this variation would suggest subtypes. Thus, if the heterogeneity for language impairment is solely individual difference then there will be little to no predictability and no distinct subtypes in a population. In contrast, systematic variation comes from consistent causal mechanisms and results in predictable patterns of clusters around predictable (at least statistically and, perhaps phenotypically) variables. Therefore, systematic variation could yield subtypes even within a broadly heterogeneous population.

Subtyping in Language Impairment Research

Researchers have long argued about the potential nature of subtypes of language impairment. Several theoretical subtype models have been proposed, primarily based upon differences in the primary parameters of language such as vocabulary and morphology and/or receptive and expressive features (e.g., Bishop 1998; Botting & Conti-Ramsden, 2004; Rapin & Allen 1983; Tomblin & Zhang, 1999). However some argue against specific subtypes and posit that the overall heterogeneity will not aggregate around predictable features (Leonard, 2014).

There are several approaches to testing for subtypes. For example, previous research has employed pre-specified language parameters (e.g., cut-off scores) to create groups (e.g., Bishop, 1997b; Tomblin et al., 1997). For example, Tomblin, Records, and Zhang (1996) categorized children as language impaired if they were below -1.25 z-scores on two or more language domains; additionally a cognitive standard score of less than 87 but greater than 70 further split the sample into children with "normal" or "low-normal" cognition. Other researchers have utilized checklists or criterion referenced experimental language tasks to divide groups into subtypes (e.g., Bishop, 1998). However, a potential limitation to cut-off score or criterion based approaches, is that cut-off scores and criterion referenced tasked (with +/- scoring) can mask variability in data, especially when the groups are arbitrarily created. An alternative is to use a more sophisticated statistical method, which employs continuous scores to explore the underlying structure in the data (Whiteman & Loken, 2006).

Continuous variables used to examine subtypes are employed with statistical techniques such as cluster analysis. There is a limited research literature that explores subtypes of language impairment using clustering techniques of this nature (Conti-Ramsden & Botting, 1999; Tomblin & Zhang, 1999), but the approach has long been used extensively in other related psychological conditions such as schizophrenia (e.g., Heinrichs & Awad, 1993) and autism (e.g., Prior et al., 1998). Clustering techniques allow for researchers to examine underlying behavioral phenotypes in the data. These techniques create groupings of cases by maximizing similarities within groupings, while also maximizing dissimilarities between groupings. However, the research using clustering techniques to date has either pre-specified the number of clusters (e.g., Conti-Ramsden, Crutchley, & Botting, 1997) or been inconclusive (Tomblin & Zhang, 1999). Pre-specifying a cluster solution is potentially problematic because the data are forced into a specified structure that may or may not match the true underlying structure of the data. Modern cluster analysis often includes two steps: 1) testing a database for a optimal number of clusters without setting an *a priori* set and 2) *post hoc* mapping statistical cluster solutions onto clinically relevant parameters to the statistical cluster solutions.

Prior research has appropriately focused largely on language measures such as vocabulary and grammar to identify subtypes within language impairment. However, differences in cognitive abilities could explain subtypes as well. Cognition is a multifaceted domain that encompasses thinking and reasoning about the world. A widely used multifaceted theory of cognition is the Cattell-Horn-Carroll theory (Carroll, 2003). Including cognitive parameters might provide more information about how children with language impairment perform broadly. Research indicates that children with language impairment have different cognitive strength and weaknesses compared to their typical peers (Botting, 2005; Swisher & Plante, 1993; Swisher, Plante, & Lowell, 1994; Tomblin & Nippold, 2014). This project considers the possibility that cognitive abilities interact with language abilities to differences in children with language impairments that might line up with subtypes of language impairment. Therefore, a cluster analysis testing for subtypes of specific language impairment (SLI) should include measures of key language domains and an estimate of g that is relatively independent of language measures.

The EpiSLI Database

In 1994, Tomblin and colleagues conducted an epidemiological study of the prevalence of language impairment (Tomblin, 2010; Tomblin et al., 1996; Tomblin et al., 1997). Prior to Tomblin and colleagues, the estimates for the prevalence of language impairment were debated. There was a need for an epidemiological sample to estimate the prevalence of language impairment more accurately (Tomblin et al., 1997). Although more than two decades old, the database, and findings arising from it continue to be foundational to language impairment research and it remains as the single largest population sample of SLI in the literature (cf. Tomblin & Nippold, 2014). This study uses the Epidemiological Study of Specific Language Impairment (EpiSLI; Tomblin, 2010) database to explore possible subtypes of language abilities in kindergartens.

The EpiSLI has language and cognitive data for nearly two thousand kindergarteners. The language measures include subtests from *Test of Language Development Primary 2nd edition* (TOLD-P:2; Newcomer & Hammill, 1988) and a narrative task developed by Culatta, Page, and Ellis (1983). The TOLD-P:2 has six subtests that tap the majority of language domains: Picture Vocabulary, Oral Vocabulary, Grammatical Completion, Grammatical Understanding, Sentence Imitation, and Word Articulation. The narrative task yields two scores: narrative comprehension and narrative retell. There are two non-verbal cognitive measures from the *WPPSI-R* (Weschler, 1989). The cognitive subtests are: Block Design and Picture Completion. Additional information is provided in the Methods section about the database and the variables.

Study Rationale

This study included kindergarteners, because the majority of children with language impairment are identified and receiving services by kindergarten and because SLI can be differentially diagnosed from other conditions such as autism and intellectual disability with high sensitivity and specificity in Kindergarten. There is a need to expand on previous research by examining subtypes by a) using continuous data rather than cut-off scores to identify subgroups and b) including estimates of g in the subtype analyses. Both of these approaches are refinements of previous research, which has historically used cut-off scores (e.g., Bishop, 2004), and/or excluded estimates of g while considered language variables (e.g., Botting & Conti-Ramsden, 2004; Tomblin & Zhang, 1999).

Cluster Analysis

Cluster analysis is a set of statistical procedures that examines how cases are interrelated (Cramer, 2003). Clusters are defined as groupings that have underlying systematic similarities (Jain, 2009; Jain, Murty, & Flynn, 1999). This methodological approach allows for the development and examination of empirically driven groupings. There are two basic types of cluster techniques, hierarchical and partitional and there are a number of statistical methods for both types (Jain et al., 1999). Comprehensive subtype analyses should use a combination of hierarchical and partitional clustering methods, because replication and convergence of subtype solutions across methods would provide additional evidence of internal consistency in the subtypes.

Some subtypes models assume that all children belong to one superordinate grouping, specifically the population of impaired language learners (Leonard, 1989), while other models assume that children with language impairment form unique subgroups (Conti-Ramsden & Botting, 1999). By using both clustering methods it is possible to investigate whether or not a superordinate category and/or identifiable subgroups are evident in the analyses. Each type (hierarchical and partitional) has advantages and disadvantages, for example, hierarchical methods are inflexible when creating clusters, but partitional methods are highly flexible. In contrast, partitional methods generally require specification of an *a priori* number of clusters, but hierarchical methods do not. Using both methods, and comparing the outcomes, will increase the strength and generalizability of the results.

Hierarchical clustering. Hierarchical clustering methods apply a variety of different formulas to generate clusters. Hierarchical clustering assumes there is a superordinate category, but with unique subgroups that can be detected using probabilistic analyses (rather than *a priori* behavioral/phenotypic analyses). Therefore one would use a hierarchical clustering method if it were assumed that is that all of the cases are somehow related. An example of this in botany would be subtypes of flowers such as irises or orchids. Each iris has more in common with member of its own subspecies, but still is subordinate to the overarching class of irises. Most methods use chaining (i.e., connecting cases) to determine how similar cases are to each other (Jain et al., 1999). Once two cases have been linked, those cases will always belong to the same cluster, regardless if dissimilarity increases as other cases are added to the cluster at another link. Among the most commonly used hierarchical clustering methods are Ward’s method, single linkage, and complete linkage (Jain et al., 1999). This study uses Ward’s method. Ward’s method has been used frequently in several fields to examine profiles of atypical populations, such as psychiatry (e.g., treatment patterns; Lambert, Brannan, Breda, Heflinger, & Bickman, 1998) and autism (Stevens et al., 2000).

Ward’s method starts with each case as a single cluster and uses minimum variance criterion to combine similar cases until there is a single cluster (Cramer, 2003; Jain et al., 1999). An advantage of Ward’s method is that the user does not have to specify the number of clusters *a priori*. However, once a case has been subsumed into a cluster, it does not change clusters; even if ultimately the case is in a cluster where the within cluster sum of squared errors is high, because there is not enough similarity in the cases (Cramer, 2003; Jain et al., 1999).

Partitional clustering. Partitional method also applies a number of different formulas. But, partitional clustering assumes that every cluster is *unique* and does *not* belong to a superordinate category. Partitional clustering techniques formulate a “prototype” case(s), called centroid(s). For partitional clustering, the cases move around the centroids until membership is stable (i.e., the similarity is high within cluster; Jain et al., 1999). The most commonly used method is K-means (Jain et al., 1999). This study will employ K-means as the partitional method, because K-means has been used extensively in behavioral research. For example, language researchers used K-means to explore subtypes of specific language impairment (Conti-Ramsden & Botting, 1999; Conti-Ramsden et al., 1997).

K-means was one of the first clustering methods to be designed and has been used for over 50 years (Jain, 2009; Steinley, 2006). Traditionally for K-means, the user has to specify the number of clusters at the beginning; however, it is possible to examine a limited number of multiple cluster possibilities. A potential threat to the approach is that K-means can converge on local optima. Applying Monte Carlo techniques (e.g., bootstrapping) and multiple starting points (e.g., different possible cluster solutions; Steinley, 2006) help prevent convergence on local optima (and spurious subtype solutions). The primary advantage and rational for using K-means to test the question of subtypes in SLI research is the flexibility in optimizing the clustering solution(s).

A Priori Subtype Models in SLI

Several subtype models exist for language impairment (e.g., Bishop, 1987; Botting & Conti-Ramsden, 2004; Rapin & Allen, 1983; Tomblin et al., 1997). Researchers developed these models based on clinical impressions and/or theoretical frameworks. Most models consider that nonverbal intelligence must be within the broad range of typical (e.g., above an "IQ" of 85) for children to be classified as language impaired (e.g., *DSM-5*). Others argue that the criterion is simply "no history of cognitive impairment" (Cole & Fey, 1996; Tomblin et al., 1997), which implies a nonverbal "IQ" of no less than 70. Because of this, subtyping should include estimates of cognitive level as these potential cognitive "subtypes" could yield clusters within the superordinate group of SLI cases.

For analysis and interpretation, this project will examine four models that are exemplars of current theoretical approaches to classifying SLI. Bishop and Tomblin each have proposed subtypes of language impairment that represents systematic variations in the deficits of children with language impairment. In contrast, Leonard and the *DSM-5* propose that there are no subtypes of language impairment, instead only individual differences, usually dependent on the severity of symptoms without predictable clusters. With regard to cognitive level, Bishop, Leonard, and *DSM-5* assume that IQ is above 85 in children with language impairment. Tomblin split his subtypes based on cognitive status, segregating those with scores above and below 87. These four models represent the most widely used models in language impairment research and diagnosis. Although, these four models are canonical in the literature of language impairment, there is no study comparing these models empirically.

DSM-5 model. The *DSM-5* (APA, 2013) defines a language disorder as a persistent deficit in acquisition and use of language in the absence of a known etiology. Furthermore, broad cognitive impairment is an exclusionary criterion. The current edition has no subtypes of language disorder (*DSM-5*) and has created separate diagnostic categories for language disorder and social communication disorder. The diagnostic features of a language disorder focus on vocabulary and grammar; specifically reduced vocabulary and limited sentence length. Additionally, the *DSM-5* indicates that deficits in discourse are linked to deficits in vocabulary and grammar. The *DSM-5* criteria allude to difficulties in syntax and pragmatics, given the emphasis on difficulties with connecting sentences or creating a narrative.

To support the *DSM-5* model, analyses would yield a cluster or a set of non-distinct clusters. The key features of DSM-5 can be examined with the variables available in EpiSLI database. Deficits would be present on the vocabulary, grammar, and narrative variables. However, Word Articulation (speech) should be within normal limits. Further, Block Design and Picture Completion, estimates of g, should not cluster the data into distinct groups. See Table 1 for the predicted performances for the cluster solution for the *DSM-5*. There could potentially be two clusters, a typical cluster and a cluster wherein all language variables are below average, but one would posit that no interpretable three, four, or five (or greater) solution could be derived from the data.

Table 1

Predicted performance for DSM-5 clusters on EpiSLI variables in z-scores

	Subtypes	
Variable	Typical	Language disorder
Picture Vocabulary	0 to +2	< -1
Oral Vocabulary	0 to +2	< -1
Grammatical Completion	0 to +2	< -1
Grammatical Understanding	0 to +2	< -1
Sentence Imitation	0 to +2	< -1
Word Articulation	0 to +2	< -1
Narrative recall	0 to +2	< -1
Narrative comprehension	0 to +2	< -1
Block Design	0 to +2	< -1
Picture Completion	0 to +2	< -1

Leonard model. Leonard argues that there are no subtypes of language impairment (Leonard, 2000, 2009, 2010). He holds that language impairment is a disruption in language learning due to distorted processing of input (Leonard, 1989). Leonard proposed that there are many domains of language that can be affected, but not every domain must be affected for a child to have specific language impairment (Leonard, 2000). His proposal suggests that there are individual differences for severity within each language domain and no consistent subtypes (Leonard, 2009, 2010). Furthermore, Leonard argues against the division of children with specific language impairment into expressive and mixed expressive-receptive subtypes (Leonard, 2009).

Under Leonard's model, individual differences in the severity of affected language domains causes the heterogeneity in children with language impairment. Furthermore, these individual differences are not principled or causal and do not lead to distinct subtypes. However, contradictory research (Conti-Ramsden & Botting, 1999) indicates that there are distinct subtypes unrelated to random individual differences in the degree of severity of language impairment. Leonard's model aggregates strengths and needs into a simple linear continuum.

To support Leonard's model, analyses would yield either a single cluster or a set of non-distinct clusters. If there were a set of non-distinct clusters, it is expected that the differences would be random. Leonard (2014) notes that deficits can be seen in vocabulary, syntax, pragmatics, grammar, and/or phonology. Using the language variables in the EpiSLI database, we can examine vocabulary, grammar, phonology, and pragmatics. Word Articulation (speech) may be within normal limits. Further, Block Design and Picture Completion should not separate the language-impaired cluster(s) into distinct groups. See Table 2 for how each EpiSLI variable taps domains important to Leonard.

Table 2

Language domains important to Leonard accounted for by the EpiSLI variables

	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	*					

Bishop model. Bishop based her first subtype model on clinical impressions. This model had five subtypes language impairment: (a) developmental apraxia of speech, (b) developmental language delay, (c) phonologic-syntactic syndrome, (d) semantic-pragmatic disorder, and (e) auditory imperceptions (Bishop & Rosenbloom, 1987). Bishop has continued to develop this subtype model over subsequent decades (Bishop, 1997b, 1998, 2004; Bishop, Bright, James, Bishop, Van der Lely, 2000; Bishop, Chan, Adams, Hartley, Weir, 2000). She has examined pertinent aspects of certain subtypes and proposed new subtypes. For example, Bishop examined the characteristics of semantic-pragmatic disorder (Bishop, 1997a; Bishop & Rosenbloom, 1987), and subsequently argued that semantic-pragmatic disorder should be reframed as pragmatic language impairment (Bishop, 1998, 2004; Bishop, Chan, et al., 2000). Bishop explored the possibility of a grammatical specific language impairment subtype; however, her evidence did not support a distinct grammatical subtype (Bishop, Bright, et al., 2000).

More recently, Bishop (2004) proposed a four-subtype model. The subtypes were: (a) typical specific language impairment (formerly phonological-syntactic), (b) severe receptive language disorder (formerly verbal auditory agnosia or pure receptive disorder), (c) developmental verbal dyspraxia, and (d) pragmatic language disorder. Typical specific language impairment is described as deficits in grammar, phonology, syntax, and semantics. Severe receptive language disorder is posited as a very rare subtype and is described as deficits in comprehension for all domains of language; children may also present with poor articulation. Developmental verbal dyspraxia is described as deficits in phonology, specifically manipulating sounds and forming representations, and poor receptive language. Pragmatic language impairment is described as deficits in social language, such as disjointed connected speech or inappropriate conversational skills; children may also present with poor grammar or syntax. Bishop has found significant differences between children with typical specific language impairment and pragmatic language impairment for skills (e.g., coherence during conversation $d = 0.94$; Bishop, 1998).

Bishop applied pre-determined classifications in her research. Thus, these *a priori* classifications make it difficult to externally determine the validity and reliability of her subtypes. However, aspects of these parameter clusters (and subtypes) can theoretically be examined within the EpiSLI database. To support Bishop's subtype, we would need to observe four clusters that have significantly below average language scores. A "typical specific language impairment" cluster would present with deficits on grammar and narrative variables. Word Articulation and vocabulary variables should be within normal limits. A "severe receptive language disorder" cluster would present with deficits on receptive variables (e.g., Picture Vocabulary, Grammatical Understanding, and narrative comprehension) greater than deficits on expressive variables (e.g., Oral Vocabulary, Grammatical Completion, and narrative recall). Word Articulation might be within normal limits. A "developmental verbal dyspraxia" cluster would present with deficits on Sentence Imitation and Word Articulation; there would also be increased deficits in the receptive variables compared to the expressive variables. A “pragmatic language impairment” cluster would present with deficits on the narrative and grammar variables. The Block Design and Picture Completion estimates of g should not be variables that separate clusters; however it is difficult to make predictions for the cognitive variables, as Bishop has not used stable cognitive criteria in her studies. See Table 3 for predicted performance for each Bishop subtype on the EpiSLI variables in z-score format.

Tomblin model. Tomblin identified two subtypes of language impairment during his epidemiological study of the prevalence of language impairment (Tomblin et al., 1997). The two subtypes of children, specific language impairment (SLI) and nonspecific language impairment (NLI), displayed deficits in language skills, yet had different nonverbal intelligence abilities. Tomblin employed standardized test scores to identify subtypes and cut-off scores (e.g., above and below 87 to separate SLI and NLI subtypes). Tomblin and colleagues have continued to compare these two language impaired subtypes (e.g., Catts, Fey, Tomblin, & Zhang, 2002).

Subsequent research has identified differences and similarities among Tomblin’s SLI and NLI subtypes of language impairment. The two subtypes exhibited differences on several language domains. For grammar, children with specific language impairment were better at tense marking verbs overall, and especially for regular past and third person *–s* at age five (*d*s = 0.29, 0.25 and 0.31 respectively; Rice, Tomblin, Hoffman, Richman, & Marquis, 2004). For syntax, children with specific language impairment had a more developed syntactic system at ages 13 (*d* = 0.77; Nippold et al., 2008) and 15 (*d* = 0.55; Nippold, Mansfield, Billow, & Tomblin, 2009). These findings suggest that children with specific language impairment have different linguistic systems than children with nonspecific language impairment. Furthermore, there were differences for long-term outcomes.

Table 3

Predicted performance for Bishop subtypes on EpiSLI variables in z-scores

	Subtype				
Variable	Typical	Typical SLI	Developmental verbal dyspraxia	Severe receptive language disorder	Pragmatic language impairment
Picture Vocabulary	0 to +2	0 to -1	< -1	< -2	0 to -1
Oral Vocabulary	0 to +2	0 to -1	0 to -1	< -1	0 to -1
Grammatical Completion	0 to +2	< -1	0 to -1	< -1	< -1
Grammatical Understanding	0 to +2	< -1	< -1	< -2	< -1
Sentence Imitation	0 to +2	< -1	< -2	< -1	< -1
Word Articulation	0 to +2	0 to -1	< -2	0 to -1	0 to -1
Narrative recall	0 to +2	< -1	0 to -1	< -1	< -2
Narrative comprehension	0 to +2	< -1	< -1	< -2	< -2
Block Design	0 to +2	0 to +2	0 to +2	0 to +2	0 to +2
Picture Completion	0 to +2	0 to +2	0 to +2	0 to +2	0 to +2

Notes. SLI = specific language impairment.

For example, children with nonspecific language impairment had significantly lower reading achievement scores (*d* = 0.52 to 0.63), and they were more likely to have a reading disability despite similar phonological awareness abilities than children with specific language impairment (Catts et al., 2002). Children with nonspecific language impairment were also more likely to maintain their language-impaired status in subsequent years (Cramer's ϕ between 0.45 and 0.66; Tomblin, Zhang, Buckwalter, & O'Brien, 2003) than children with specific language impairment. These findings suggest different developmental trajectories for the two subtypes. There were also differences associated with nonverbal skills, such as spatial working memory, where children with specific language impairment outperform children with nonspecific language impairment (*d* = 0.60; Karasinski & Ellis Weismer, 2010). These nonverbal differences suggest that there may be broader cognitive differences between the two subtypes. However, there were similarities between the subtypes, which might suggest that the two subtypes do not have meaningful clinical differences. For language abilities, there were no differences for tense marking on irregular past tense verbs (*d*s between 0.15 and -0.09; Rice et al., 2004) or non-word repetition (Ellis Weismer et al., 2000). And there were no differences for speed of processing for motor or cognitive tasks (Miller et al., 2006).

To support Tomblin's two-subtype model, we would need to find two distinct clusters with below average language scores. See Table 4 for how children with SLI and NLI performed on the language and cognitive variables from the EpiSLI database. There are effect sizes ranging from medium (*d* = 0.48 for Grammatical Understanding) to very small (*d* = 0.09 for narrative recall). These effect sizes relate to previously differences between the subtypes from the longitudinal research. The clusters should have differences for grammar variables (receptive and expressive), and might be different for vocabulary variables (either receptive or expressive). Previous work by Tomblin and Zhang (1999) suggests that Word Articulation is a language domain governed by other cognitive abilities, so it is likely there will be no difference for this variable. It is also predicted that the narrative variables will either be the same or only slightly differentiated. Block Design and Picture Completion differentiate two clusters.

Table 4

Performance on EpiSLI variables for language impaired subtypes from Tomblin

	Subtype		
Variable	SLI	NLI	*d*
Picture Vocabulary	8.51 (2.26)	7.67 (2.47)	0.36
Oral Vocabulary	7.48 (1.80)	7.14 (1.57)	0.20
Grammatical Completion	7.45 (1.32)	7.06 (1.48)	0.28
Grammatical Understanding	6.55 (2.07)	5.61 (1.86)	0.48
Sentence Imitation	5.72 (1.73)	5.39 (1.59)	0.19
Word Articulation	8.56 (3.61)	7.88 (3.47)	0.19
Narrative recall[a]	3.70 (2.81)	3.47 (2.59)	0.09
Narrative comprehension[a]	5.24 (1.79)	4.95 (1.89)	0.16
Block Design	9.59 (2.24)	5.34 (1.99)	2.01
Picture Completion	10.51 (2.22)	6.43 (2.14)	1.87

Notes. Mean (sd). *d* = Cohen's d. SLI = specific language impairment, NLI = nonspecific language impairment.

[a] Reported as raw scores, all others are scaled scores.

Research Aims

The research literature on subtypes of language impairment is inconclusive. There are several competing models of SLI, both with and without distinct subtypes. There are several limitations to the previous research, such as employing on cut-off scores and/or criterion reference measures as well as lack of generalization or replication. The overall heterogeneity in the SLI research provides reason to suspect that there are systematic variations that give rise to subtypes and there are clinical models suggesting assessment and intervention ramifications for subtyping. If there are indeed subtypes, much can be gained by understanding how these subtypes perform, both for researchers and clinicians. This study addresses the following key gaps in the research on subtypes: (a) applying continuous variables without pre-specifying the number of clusters, (b) including estimates of g (cognitive ability) to expand the understanding of possible subtypes, and (c) comparing models.

Aim 1. The first aim explores the nature and extent of clustering that occurs in the EpiSLI database without an *a priori* specification of the number of clusters. The validity of current SLI theories can be examined by testing for empirically derived clusters of language abilities of kindergarteners. This first aim is exploratory. Therefore no *a priori* hypothesis is provided for the number of cluster or how the clusters will perform. The purpose of this exploratory analysis is to empirically determine the number and characteristics of the possible clusters

Aim 2. The second aim compares the empirically derived clusters to the four aforementioned theories. This comparison will be done by statistically comparing the empirical results to either reported data or pre-specified patterns of performance to determine whether or not the clusters derived from the analyses support those models. Although these four models are used in clinical practice and research, there has been limited comparison of these models using probabilistic (assumption free) analyses. Such a comparison and examination would inform clinical and research practice. Because the current SLI models differ regarding whether there are subtypes (and associated clusters) a two or four cluster solution could support either Tomblin's or Bishop's model (respectively) might be supported, whereas both Leonard and the *DSM-5* will be supported if there are no subtypes (clusters) detected.

CHAPTER 2

METHODS

This study used a publically accessible database, EpiSLI (Tomblin, 2010), to address the study aims. This study analyzed the EpiSLI database for clusters within the data to determine if those clusters represent unique subtypes of language ability in kindergarteners. Further, if clusters were detected, the empirically derived clusters would be compared to two current models of subtypes derived from the current language impairment research literature. If no clusters were derived, two models of SLI that directly (Leonard) or indirectly (DSM-5) reject subtypes would be supported. This study employed two types of clustering techniques, hierarchical and partitional to converge on an appropriate number of clusters (Ward's and K-Means). Descriptive information was retrieved to examine the profiles of the subtypes if warranted.

EpiSLI Database

Overview. This project utilized the EpiSLI database (Tomblin, 2010). The EpiSLI database was shared with the PI (Hope Lancaster) under the National Institute for Deafness and Other Communication Disorders (NIDCD) public sharing policy. For more information on the EpiSLI datashare see Tomblin (2010).

Tomblin et al. (1997) conducted a stratified-epidemiological study in Iowa testing children in urban, suburban, and rural counties to estimate the prevalence of language impairment. Tomblin et al. (1997) screened an epidemiological sample for language impairment. Children who failed the screening and a random thirty-three percent of children who passed the screening completed a comprehensive diagnostic battery for language, cognition, hearing, and other skills. The EpiSLI database contains language, cognition, and physiological (e.g., gait, audiometric) measures for 1,929 subjects. Study design, data collection, variables, and other information are summarized in Tomblin and Nippold (2014).

Subjects. All subjects were between the ages of 5 and 6. There are 1,929 subjects, spilt into four discrete groups (Typically Developing n = 1226; Low Normal n = 198; Specific Language Impairment n = 277; Nonspecific Language Impairment n = 228). There is no difference between the groups on age (M = 6 years; SD = 4 months) or gender (between 52 and 59 percent male). The EpiSLI database was designed to be representative of the population based on the 1990 US census data. The 1990 US census reported that the population was 80 percent White, 12 percent Black, and 7.5 percent other minorities. The EpiSLI sample had 84 percent White, 13 percent Black, and 2.7 percent other minorities, which matched the 1990 census for White and Black, but underrepresented other minorities. See Table 5 for descriptive variables for all four groups.

Table 5

Descriptive breakdown of EpiSLI database

Subgroup	N	Age	Gender	Language	Cognition	Race
Typically Developing	1226	6 (3.8)	54%	-0.08 (0.67)	0.39 (0.62)	89.3% / 8.6%
Low Normal	198	6 (3.7)	53%	-0.55 (0.51)	-1.09 (0.38)	72.2% / 22.2%
SLI	277	6 (3.7)	59%	-1.52 (0.38)	0.12 (0.53)	84.8% / 12.3%
NLI	228	6 (3.8)	52%	-1.76 (0.48)	-1.27 (0.48)	65.8% / 31.1%

Notes. Mean (SD). For Age mean is reported in years and SD in months. Gender is reported as percent male. Language is a composite of five TOLD-P:2 (Newcomer & Hammill, 1988) and two narrative variables (Culatta et al., 1983) as described in Tomblin et al. (1996). Cognition is the average of z-scores of Block Design and Picture Completion (Weschler, 1988). Race is percentage of White over percentage of Black.

Subjects completed language and cognition testing in a single session. Data collection occurred from 1993 to 1995. There were two cohorts, 1993 – 1994 and 1994 – 1995. Children were divided on composite language scores into typical and language impaired. Tomblin et al. (1996) required children have two language composite scores below a -1.25 z-score. Children were also split using cognitive scores. Children with typical language were split into “Typically Developing” and “Low Normal” based on whether or not their cognitive scores were above or below a standard score of 87. Children with language impairment were split into "SLI" and "NLI" according to the same cognitive criteria.

Variables. This project examined the language and cognition variables. There were eight language and two cognitive variables (Tomblin et al., 1997). Six of the language variables come from the *Test of Language Development – Primary* (2nd ed.; -P:2, Newcomer & Hammill, 1988), these variables were: (a) Picture Vocabulary, (b) Oral Vocabulary, (c) Grammatical Understanding, (d) Grammatical Completion, (e) Sentence Imitation, and (f) Word Articulation. Two language variables were derived from a narrative task Culatta et al. (1983) developed: (a) narrative recall and (b) narrative comprehension. The cognitive variables, Block Design and Picture Competition, came from *Weschler Preschool and Primary Scales of Intelligence Revised* performance scale (WPPSI-R; Weschler, 1989).

Tomblin used the TOLD-P:2 (Newcomer & 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, 1997a; Leonard, 2000; Newcomer & Hammill, 1988). Narrative tasks were included because narrative performance is able to predict academic performance above and beyond other language abilities (Feagans & Applebaum, 1986; Tomblin et al., 1996).

This project combined language and cognitive variables because language and cognition are both necessary to create detailed profiles of strengths and needs of children with LI. The eight language variables were included in this project because, in existing theoretical models, language variables similar to those in Tomblin et al. (1997) are often employed to define subtypes. The two cognitive measures were introduced for two reasons: (a) Tomblin used these variables to determine his subtypes by separating the groups based on a cut-off score, and (b) previous research by the PI indicates that cognition has an additive value to understanding a child's academic abilities (Lancaster & Camarata, 2013). See Table 6 for variables by construct and source.

Language measures. Language variables were a combination of a standardized test battery, the TOLD:P-2 (Newcomer & Hammill, 1988), and a narrative task (Culatta et al., 1983). The TOLD:P-2 is an omnibus language measure that includes domain specific subtests. These TOLD:P-2 subtests include: Picture Vocabulary, Oral Vocabulary, Grammatical Completion, Grammatical Understanding, Sentence Imitation, and Word Articulation. The Culatta narrative task involved listening to a story, answering questions, and story recall. The narrative task yielded a narrative comprehension and a production score. The narrative comprehension score was based on the number of questions answered correctly. The narrative recall score was based on the number of propositions from the original story that children produced in their story retell.

Table 6

EpiSLI variables by domain and source

Variable	Source	Reliability
Language		
Picture Vocabulary	TOLD-P:2	.89[a]
Oral Vocabulary	TOLD-P:2	.89[a]
Grammatical Completion	TOLD-P:2	.76[a]
Grammatical Understanding	TOLD-P:2	.96[a]
Sentence Imitation	TOLD-P:2	.91[a]
Word Articulation	TOLD-P:2	.86[a]
Narrative recall	Culatta, et al. (1983)	n/a
Narrative comprehension	Culatta, et al. (1983)	n/a
Cognition		
Block Design	WPPSI-R	.80[b]
Picture Completion	WPPSI-R	.82[b]

Notes. TOLD-P:2 = *Test of Language Development – Primary 2nd edition* (Newcomer & Hammill, 1988); WPPSI-R = *Weschler Preschool and Primary Scales of Intelligence – Revised* (Weschler, 1989).

[a] Average test-Retest reliability coefficients reported by Newcomer & Hammill (1988); [b] Population estimated test-retest coefficients reported by Weschler (1989).

Item analysis validated the TOLD-P:2. 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 has some minor weaknesses as indicated by Word Articulation's low scores for internal consistency ($r = .67$ to .93) and stability ($r = .74$), and Grammatical Understanding's low stability ($r = .74$).

Cognitive abilities. Tomblin et al. (1997) employed Block Design and Picture Completion from the WPPSI-R (Weschler, 1989) to assess nonverbal intelligence. In Block Design, children use blocks to create a specified pattern in a short amount of time. Children are shown a pattern and asked to recreate that pattern using the blocks given to them as quickly as possible. For Picture Completion, children identify what is missing in a picture (e.g., an ear from a cat). Despite the fact that Picture Completion is supposed to be nonverbal, children must provide a verbal response for each item.

The WPPSI-R (Weschler, 1989) was developed to test verbal and performance cognitive ability in children. The WPPSI-R has decent validity as indicated by high concurrent validity and predictive validity (Weschler, 1989). There were no reported tests of validity for individual subtests. Weschler (1989) 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 nonverbal intelligence ($r = .73$) therefore any lack of clustering would not be a result of problems with the variables.

Data Analyses

Aim 1 explored the nature and extent of clustering in the EpiSLI database using K-means and Ward's methods. Bayesian Information Criterion (BIC) was used to determine optimal cluster size. Aim 2 examined the means and standard deviations for empirically derived clusters that were both small and stable. The analyses are described below. Appendix A provides the R syntax used for the analyses. Appendix B provides a brief tutorial on clustering and an explanation of graphs.

Data preparation. The standard scores for the variables were transformed into z-scores. This transformation insured that the variables had equal variance and were continuous. Scores were transformed using the whole database in Statistical Package for the Social Sciences (SPSS; IBM, 2012).

Four EpiSLI datasets were used for analyses. The primary two datasets were "Language Impaired" and "Typically Developing." The Language Impaired dataset contained all cases that had been originally classified as either SLI or NLI. The Language Impaired dataset was additionally spilt into SLI and NLI to conduct finer scaled analyses. The Typically Developing dataset contained cases that were classified as typically developing by Tomblin. Cases originally classified as "Low Normal" were not included, because this study was not considering how children with a potential learning disability might cluster on language variables.

Clustering. All analyses were conducted in R (2010) using the packages "cluster" (Maechler, Rousseeuw, Struyf, & Hurber, 2014) and "matrixStats" (Bengtsson, 2010). There were three steps for Ward's clustering. First, a distance matrix for the squared Euclidean distances was created using the *dist* function. Second, the data from the distance matrix was clustered using the *hclust* function. Third, Bayesian Information Criterion (BIC) was calculated for each cluster size (up to 100) to identify the cluster size corresponding to the smallest BIC value. Ward's cluster results are shown in two graphs: a dendrogram and a BIC curve.

There were three steps for K-means clustering. First, the original data were clustered using the *kmeans* function to test a range of clusters. This study tested up to 50 clusters. Second, as with the Ward's method, BIC was computed for all cluster sizes. For K-means clustering, it is always possible that a cluster solution could converge by chance. Therefore, to gauge how likely it was that the solution for a given cluster size converged by chance, a third step was included in the K-means analysis. This third step was to create random permutated data, conduct the clustering analysis, and calculate the BIC value for each possible cluster size 250 times. The permutation's code was adapted from the *kmeans* function. For the permutations, the mean and the confidence interval for the BIC for each cluster size was calculated. The BIC curves were compared, looking particularly at whether the original results were within the range of the confidence interval. In addition, the minimum BIC value for the original data was compared to the minimum mean for the permutated data. K-means clustering results are displayed in one graph showing the BIC curve for the original data and the random permutation curve with its confidence interval.

For both Ward's and K-means, the BIC was used for determining optimal cluster size. In contrast to other likelihood-based criteria, BIC has a stricter penalty for cluster size (Schwarz, 1978). See Appendix B for additional information on BIC values and their interpretation. The optimal cluster size for each method and dataset was obtained by finding the minimum value of BIC for that specific analysis. BIC was calculated for all possible clusters using within cluster sum of squares, the natural log of n, and the number of clusters. BIC functions were written for both Ward's and K-means.

A cluster had to be small and stable before it was interpreted. A small cluster size was (a) smaller than the optimal permutated cluster size and (b) less than six. The small criteria were based on theory and data interpolation. The random permutated data were not clustered by definition and represented chance clustering. Theoretically the largest possible solution considered was a five-cluster solution based on Bishop (2004). A stable cluster finding was one that did not include the random permutated cluster size in the range of optimal cluster sizes.

Principal components analysis was used to remove redundancy in the variables. Principal components analysis is a statistical procedure that transforms presumably correlated data into a set of linear and uncorrelated variables, or components. There were five steps for this study. The first step was to compute the components using the *princomp* function. The second step was to examine the components. The third step was to repeat K-means clustering, including the random permutations, with a set of components that accounted for 90 percent of the variance. The fourth step was to select the optimal cluster sizes using BIC. The fifth step was to plot the data using the first two components for both the Ward's and K-means data. There are three graphs produced from the principal component analyses: a scree plot, the BIC curves for K-means clustering, and scatter plots.

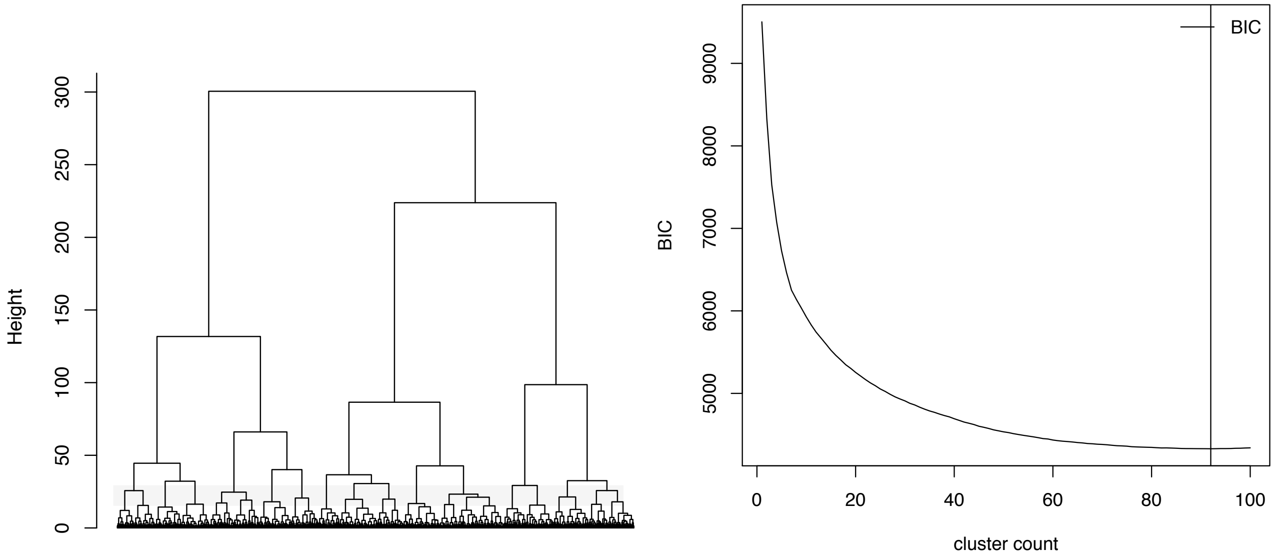
CHAPTER 3

RESULTS

The results are presented by EpiSLI dataset, which was divided by clinical status: Typically Developing (TD) and Language Impaired (LI). The LI dataset was further divided into SLI and NLI subsets. The order of presentation for the statistical analyses is as follows: Ward's clustering results are presented first; K-means and the permutated k-means results are presented second and the principal component analysis results are presented last. Within each analysis method, BIC results will be presented.

Typically Developing Dataset

The cluster size corresponding to the smallest BIC value was 92 for Ward's algorithm. Figures 1 and 2graphically represents the findings for the Ward's algorithm.



Figures 1 and 2. Ward's hierarchical clustering dendrogram (on left) and the BIC curve for clusters 1 through 100 (on right) for Typically Developing dataset. Height is the criterion value for a merge and represents the total distance accounted for by that merge. The gray box on the dendrogram and the vertical line represent the identified minimum BIC value.

The K-means algorithm converged as was demonstrated by production of an integer for all runs. Figure 3 is a graph of the BIC curves for both the original data and the permutated data. The minimum BIC was 32 clusters for the Typically Developing dataset. For the permutated data the minimum BIC was 37 clusters. The BIC values for the Typically Developing dataset were consistently smaller than chance as seen by the large gap between the two curves. Additionally the confidence interval for the permutated results does not overlap with the BIC curve for the Typically Developing dataset. The confidence interval increased as the cluster size was increased from one cluster to approximately ten clusters as is evident from visual inspection. This confidence interval increase suggested an increase in the degree of uncertainty, as cluster size got larger. The K-means results suggested that there was structure to the data, however the optimal cluster solution did not met the criteria for interpretation due to the large cluster size.

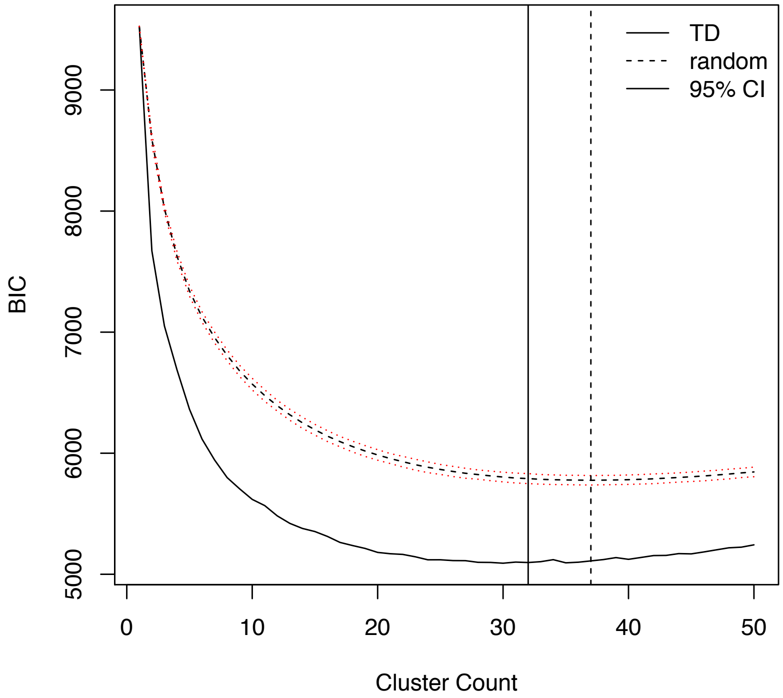


Figure 3. BIC curves for K-means clustering for Typically Developing dataset. BIC curves for original data, permutated data, and 95% confidence interval envelope. The vertical lines represent the mode cluster solution for the data and the optimal solution for the permutated data.

Principal component analysis was used to reduce redundancy in the variables by creating components that were orthogonal relationships between the language and cognitive variables. To capture 90 percent of the variance at least eight components were needed. Figure 4 shows the scree plot for the components. A scree plot graphically represents the percent variance accounted for by a component. Here component one accounted for 29 percent of the variance. Components are listed in order of decreasing variance. The K-means algorithm clustered the data with components one through eight. The PCA results showed that the overall value for BIC was reduced, although the minimum cluster sizes were larger: 32 for Typically Developing data and 36 for permutated data. Clustering with principal components decreased the gap between the BIC curves compared to the prior K-means clustering results. Figure 5 shows the BIC curves for the principal component results and the confidence interval envelope. There was less of a gap between the two BIC curves than the prior K-means clustering results. The PCA permutated data had a smaller confidence interval envelope than the permutated data using the original variables. These results suggest that using PCA eliminated much of the structure seen in the previous K-means clustering for the Typically Developing dataset as is evident from the BIC curve.

The scatterplots contain component one and two because these components account for the majority of the variance. See Figures 6 and 7 for the scatterplots. Each point is the coordinate point for a z-score on component one and component two. The circles represent “clusters” of cases. A scatterplot can be generated for any cluster solution. For ease of interpretation a two-cluster solution is presented. Scatter plots using PCA show no distinct clustering in the data, supporting the BIC results. This is true when using the distance information for Ward’s or the non-transformed data used for K-means. The clusters were non-distinct because there were not clear groupings and the circles overlap.

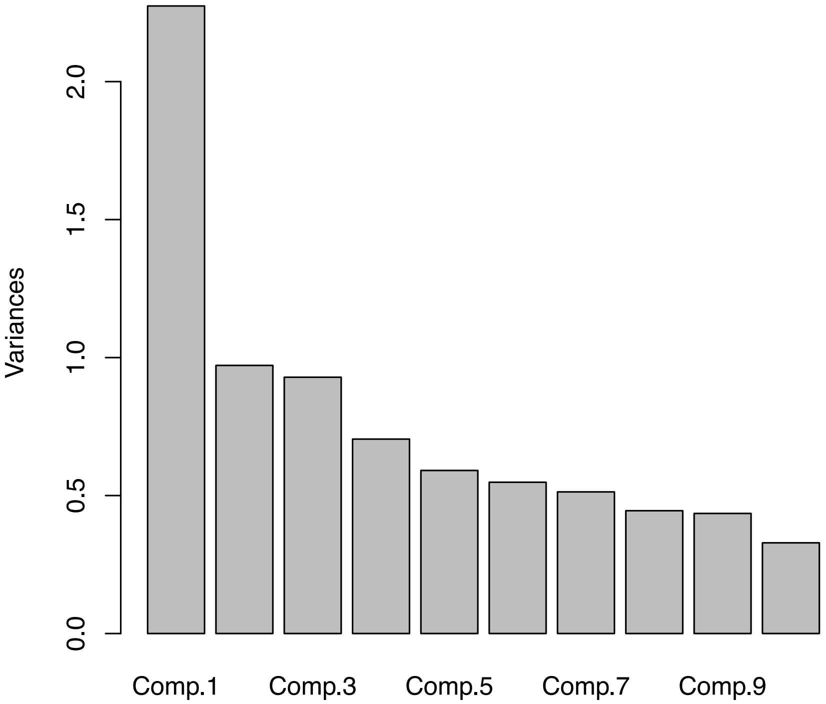


Figure 4. Scree plot of components from principal component analysis for Typically Developing dataset. Components presented in order of decreasing variance.

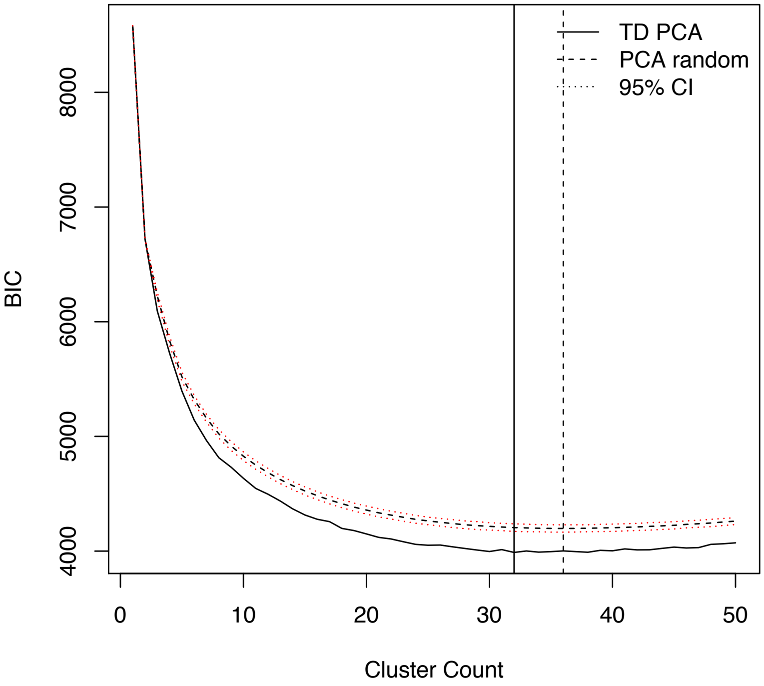
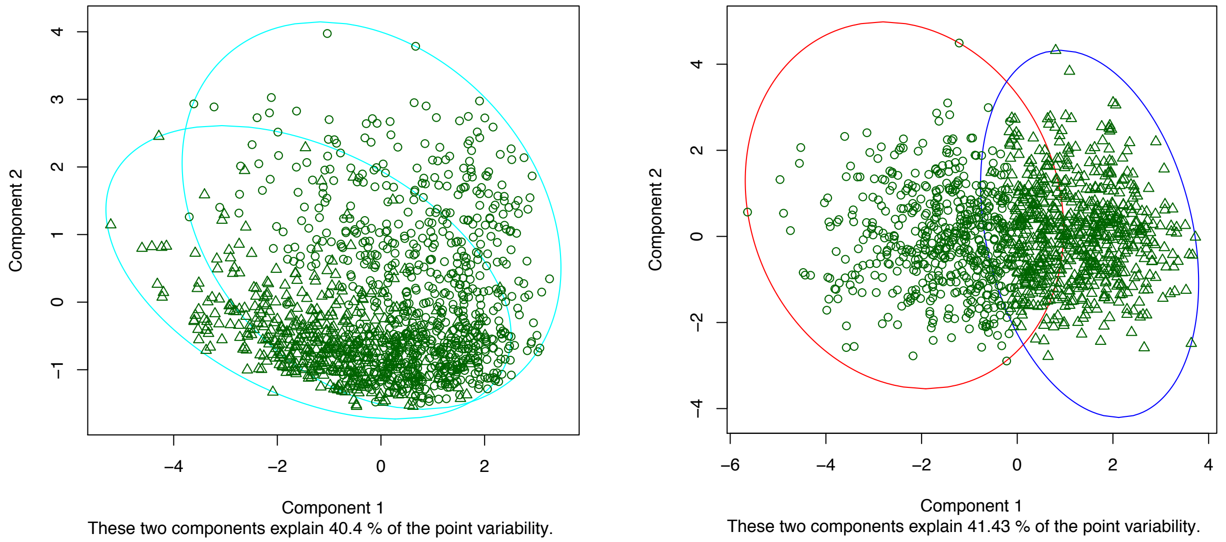


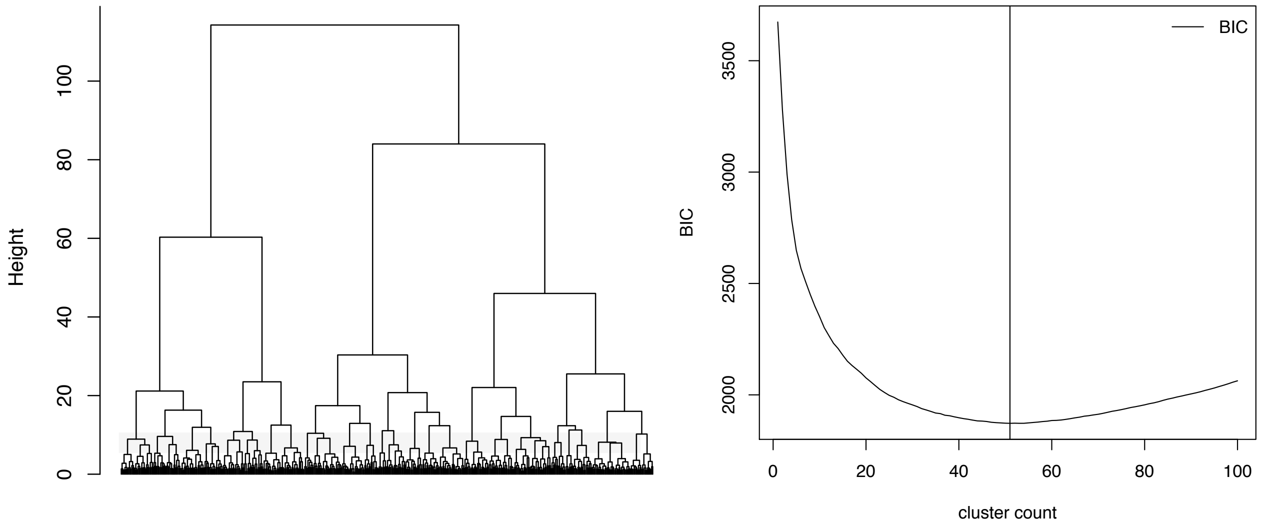
Figure 5. BIC curves when using components for Typically Developing dataset. Curves for PCA results, permutated data, and 95% confidence interval envelope for permutated data. The vertical lines represent the optimal cluster solutions.



Figures 6 and 7. Ward's cluster scatterplot (on left) and K-means cluster scatterplot (on right) for Typically Developing dataset. The circles represent clusters.

Language Impaired Dataset

The Language Impaired dataset contained cases classified by Tomblin as either SLI or NLI. This dataset provided the principal findings about language impairment subtypes for this study. The cluster size corresponding to the smallest BIC value was 51 for Ward's method. Figures 8 and 9 graphically represent the findings for the Ward's algorithm.



Figures 8 and 9. Ward's hierarchical clustering dendogram (on left) and BIC curve for clusters 1 through 100 (on right) for Language Impaired dataset. Height is the criterion value for a merge and represents the total distance accounted for by that merge. The gray box on the dendrogram and the vertical line represent the identified minimum BIC value.

K-means algorithm converged as was demonstrated by production of an integer for all runs. Figure 10 is a graph of the BIC curves for the original data, permutated data, and the confidence interval. The minimum BIC was 18 clusters for the data. The minimum BIC value was 18 clusters for the permutated data. The optimal cluster size was smaller than the optimal cluster size for the Typically Developing dataset. The BIC values were consistently smaller than chance for the Language Impaired dataset as seen by the large gap between the two curves, which was consistent with the Typically Developing dataset. Additionally, the confidence interval for the permutated results does not overlap with the BIC curve for the Language Impaired dataset, again consistent with the Typically Developing dataset. This finding supported the possibility of structure in the data. The confidence interval increased as the cluster size was increased from one cluster to approximately ten clusters as is evident from visual inspection. This increase suggested an increase in the degree of uncertainty as cluster size got larger. The K-means results suggested that there was structure to the data, however the optimal cluster solution did not met the criteria for interpretation because (1) there was a large number of clusters and (2) inclusion of the permutated minimum in the range of possible solutions.

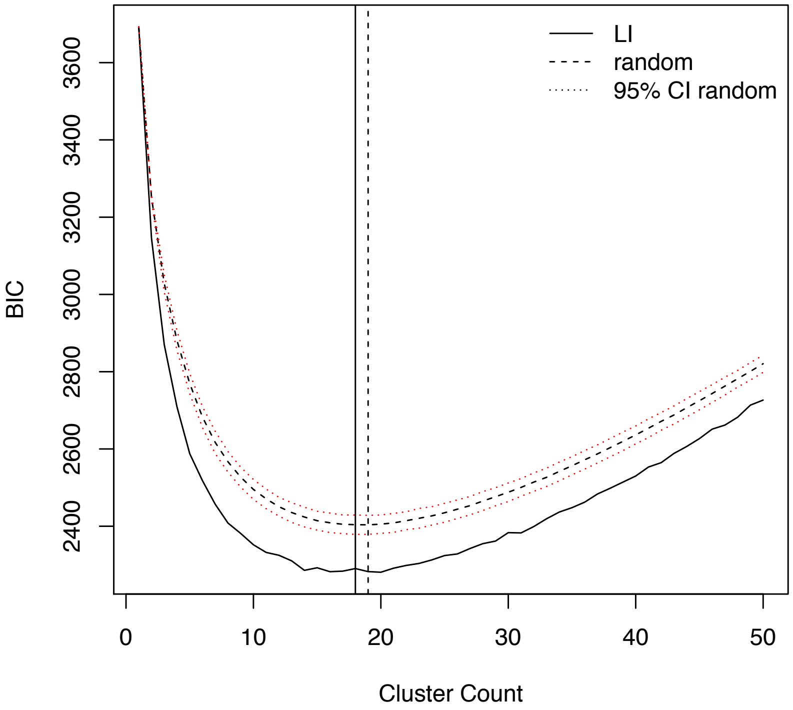


Figure 10. BIC curves for K-means clustering for Language Impaired dataset. BIC curves for original data, permutated data, and 95% confidence interval envelope for permutated curve.

Principal component analysis was used to reduce redundancy in the variables by creating components that were orthogonal relationships between the language and cognitive variables. To capture 90 percent of the variance at least eight components were needed. The first component accounted for 24 percent of the variance. Figure 11 shows the scree plot of the components. The K-means algorithm clustered the data with components one through eight. The PCA results showed that the overall value for BIC was reduced, although the minimum cluster sizes were larger: 22 clusters for Language Impaired data and 21 clusters for permutated data. Clustering with principal components decreased the gap between the BIC curves compared to the prior K-means clustering results. This finding was similar to the Typically Developing dataset. Figure 12 shows the BIC curves for the principal component results and the confidence interval envelope. The Language Impaired BIC curve overlapped with lower boundary of the confidence interval. The PCA permutated data had a larger confidence interval envelope than prior K-means clustering results. These results suggest that using PCA eliminated much of the structure seen in the previous K-means clustering for the Language Impaired dataset as is evident from the BIC curve. Figures 13 and 14 show the scatterplots for Ward's and K-means clustering data. Again, for ease of interpretation, a two-cluster solution is presented. The scatterplots showed no distinct clustering in the data, supporting the BIC results. These scatterplots were similar to the Typically Developing dataset.

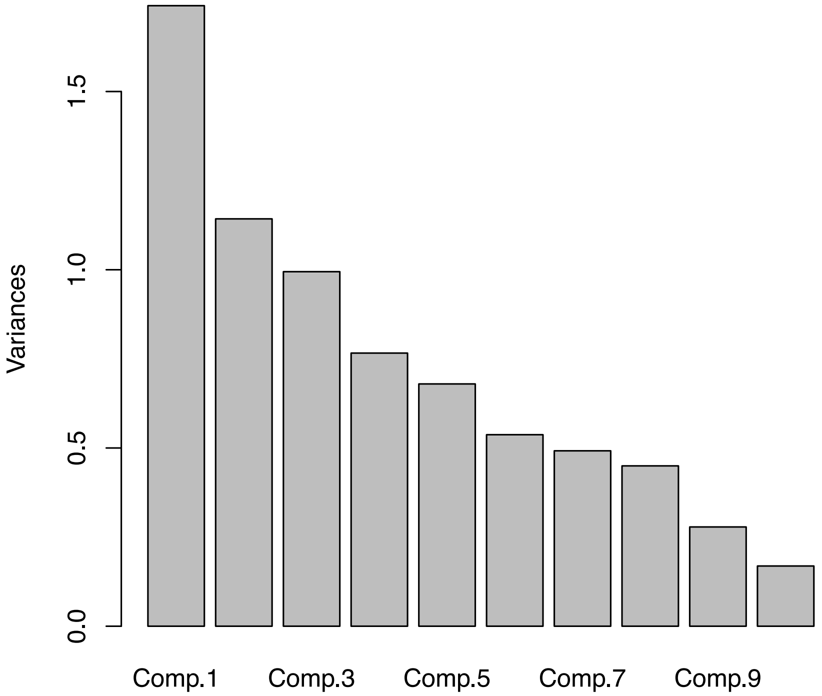


Figure 11. Scree plot of components from principal component analysis for Language Impaired dataset.

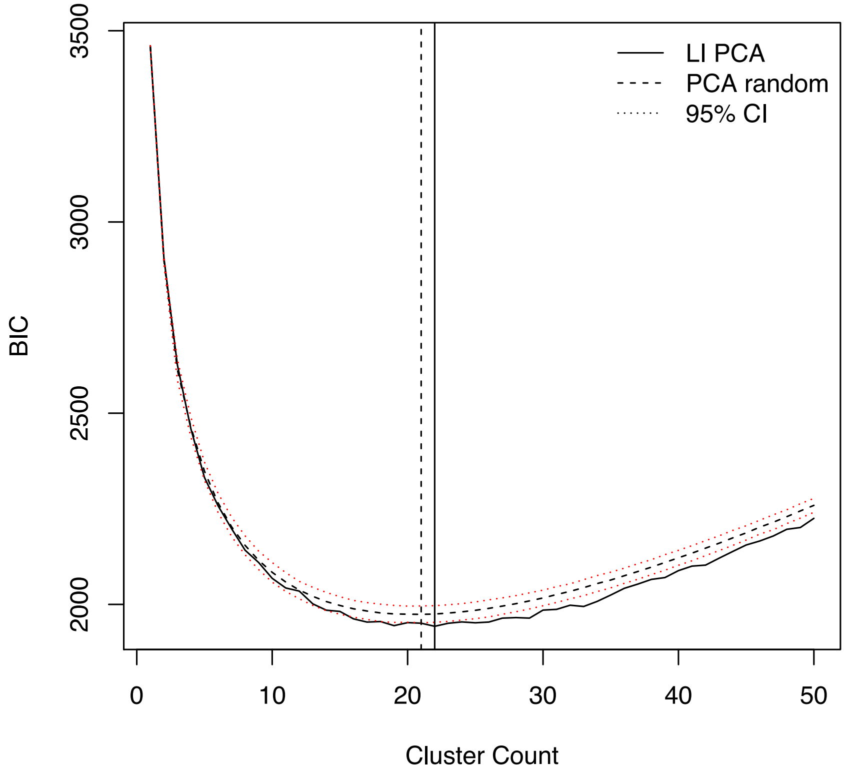
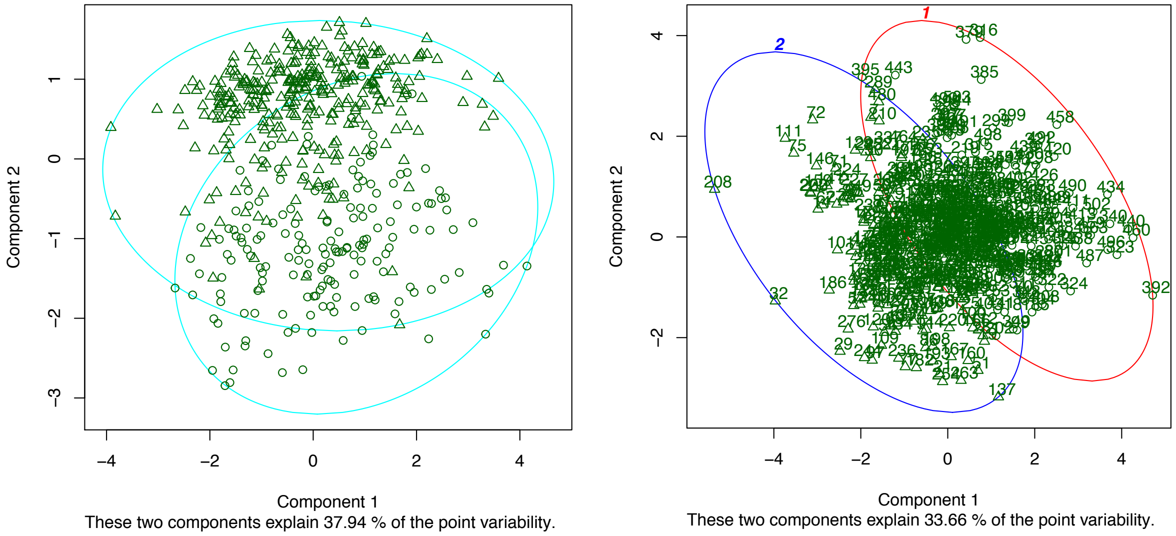


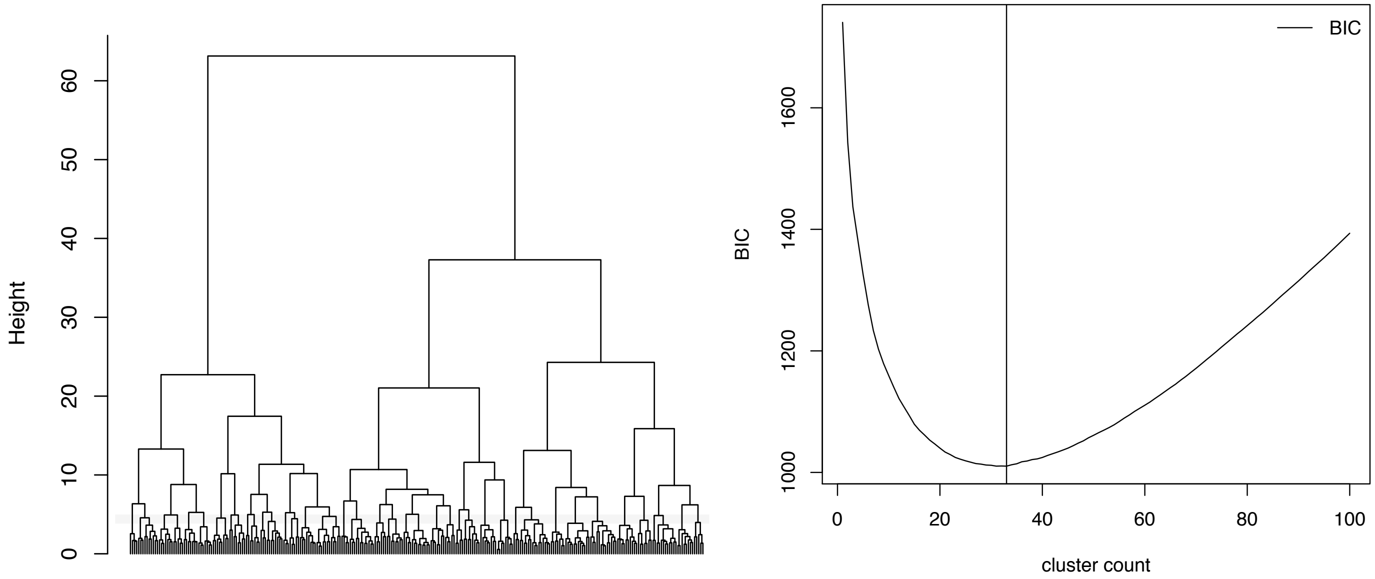
Figure 12. BIC curves when using components for Language Impaired dataset. Curves for PCA results, permutated data, and 95% confidence interval envelope for permutated curve. The vertical lines represent the minimums BIC values using principal components.



Figures 13 and 14. Ward’s cluster scatterplot (on left) and K-means cluster scatterplot (on right) for Language Impaired dataset. The circles represent clusters.

SLI only dataset. The SLI dataset was a subset of Tomblin's Language Impaired dataset, constructed by omitting the NLI participants, and was analyzed separately to generate a direct comparison to the Bishop model. The SLI dataset restricted the dataset to "normal" IQ cases as conceptualized by Bishop and thus contained *only* cases classified as SLI by Tomblin, who included children with IQ scores less than 85, but higher than 70, in his Language Impaired group. As the SLI dataset was a subset of the Language Impaired dataset, the results were compared to the larger Language Impaired dataset.

The minimum BIC value was 33 clusters for Ward's method. Figures 15 and 16 graphically represent the findings for the Ward's algorithm. The K-means algorithm converged as was demonstrated by production of an integer for all runs. Figure 17 is a graph of the BIC curves for both the original data and the permutated data.



Figures 15 and 16. Ward's hierarchical clustering dendogram (on left) and BIC curve for clusters 1 through 100 (on right) for SLI dataset. Height is the criterion value for a merge and represents the total distance accounted for by that merge. The gray box on the dendrogram and the vertical line represent the identified minimum BIC value.

The minimum BIC value was 10 clusters for the SLI dataset. The minimum BIC value was 11 clusters for the permutated data. Additionally the BIC curve for the data overlapped with the lower boundary of the confidence interval for the permutated data. This finding suggested that the SLI dataset had no structure, which contrasted with results from the larger Language Impaired dataset. The clustering results were close to chance as is evident by the occasional overlap between the SLI BIC curve and the lower bound of the confidence interval for the permutated data. The confidence interval around the BIC values increased in the permutated results as the cluster size was increased, until the cluster size reached approximately 15 at which point the confidence interval envelope began to shrink. Therefore the SLI clusters were not interpreted because the results (1) had a large number of clusters and (2) included of the permutated minimum in the range of possible solutions.

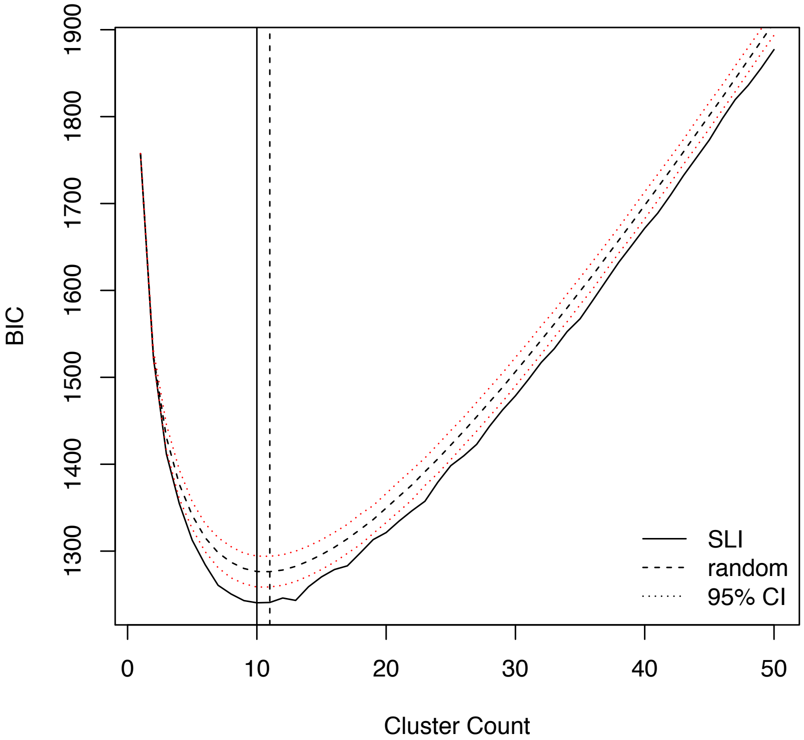


Figure 17. BIC curves for K-means clustering for SLI dataset. BIC curves for original data, permutated data, and 95% confidence interval envelope for permutated curve.

Principal component analysis was used to reduce redundancy in the variables. To capture 90 percent of the variance at least eight components were used. Figure 18 shows the scree plot of the components. The first component accounted for 22 percent of the variance. The K-means clustering algorithm clustered the data with components one through eight as the variables. The minimum BIC values were 14 clusters for SLI data and 12 clusters for permutated data. Figure 19 shows the BIC curves for the principal component results. Similar to the previous K-means, the BIC curve for the SLI original data was close to or overlapping the lower confidence interval. The confidence interval envelope increased as cluster size increased for permutated BIC curve until approximately a cluster size of 30 when the confidence interval began to decrease. The PCA replicated the prior K-means clustering results consistent for the SLI data.

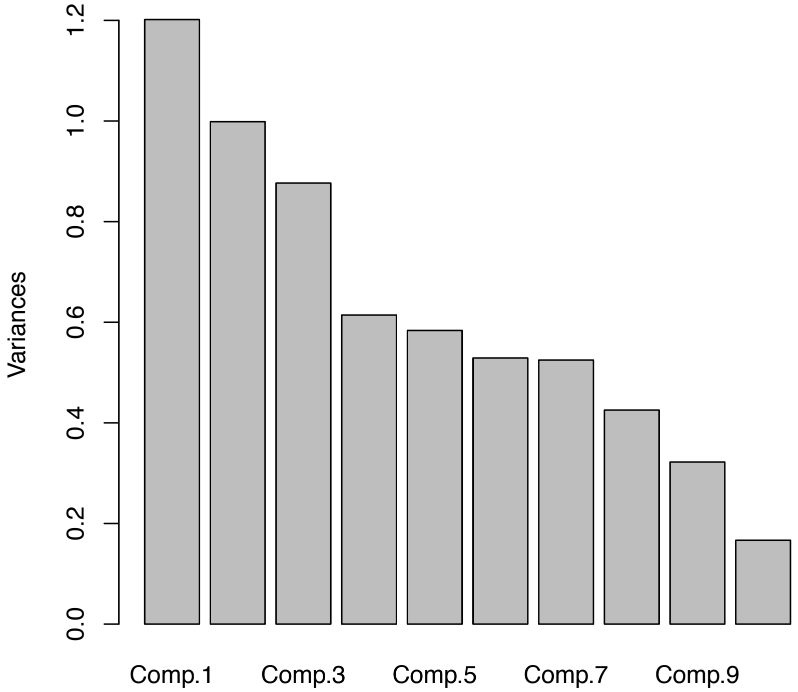


Figure 18. Scree plot of components from principal component analysis for SLI dataset.

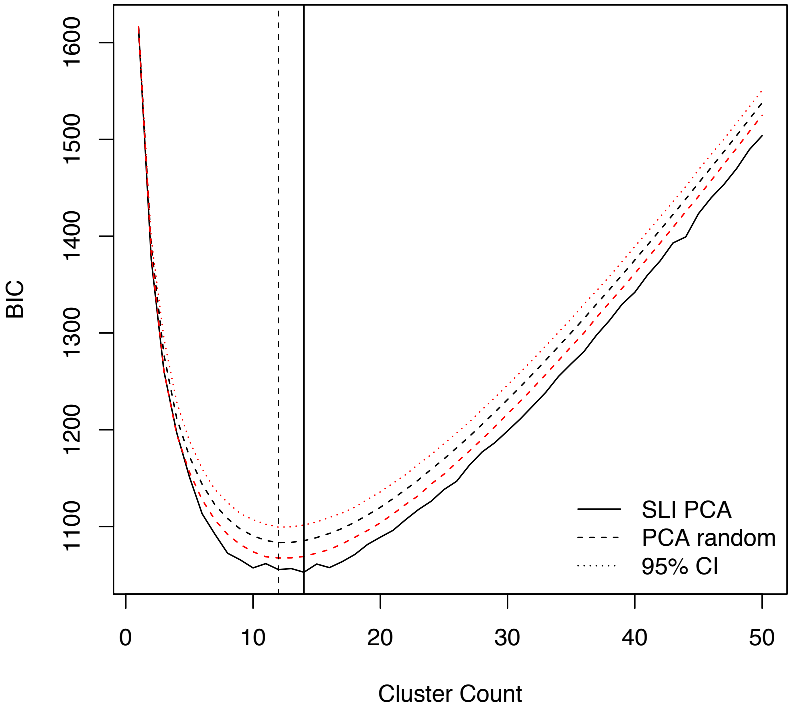
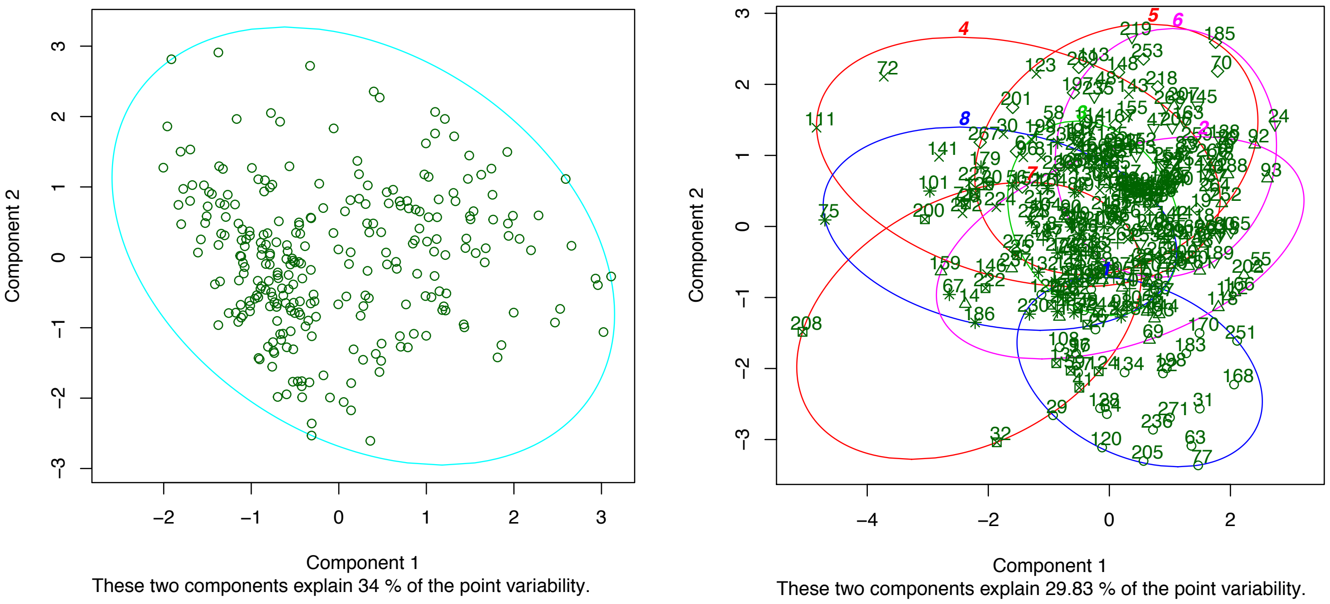


Figure 19. BIC curves when using components for SLI dataset. Curves for PCA results, permutated data, and 95% confidence interval envelope for permutated curve.

The scatterplot for Ward's method shows a one-cluster solution for ease interpretation. In contrast the second scatterplot presents a ten-cluster solution, which was the mode for the K-means results. See Figures 20 and 21 for the scatter plots. The scatterplots showed no distinct clustering in the data, similar to the BIC results and the Language Impaired scatterplots.



Figures 20 and 21. Ward's cluster scatterplot (on left) and K-means cluster scatterplot (on right) for SLI dataset.

NLI only dataset. The Language Impaired dataset was subdivided into the SLI and NLI datasets to (a) generate an SLI comparison for the Bishop model and (b) examine the NLI group, as defined by Tomblin in order to compare this dataset to the results for the SLI analyses presented above. The NLI dataset contained cases that were classified as NLI by Tomblin: children who had IQ scores below 87 and above 70 while also meeting the criteria for having significant language deficits. Because the NLI dataset was a subset of the Language Impaired dataset, the results were compared to the larger Language Impaired dataset and the SLI only dataset.

The minimum BIC value was 25 clusters for the results of the Ward's method analysis. Figures 22 and 23 graphically represents the findings for the Ward's algorithm. The K-means algorithm converged as was demonstrated by production of an integer for all runs. Figure 24 is a graph of the BIC curves for both the original data, permutated data, and a confidence interval for the permutated data. The minimum BIC value was 8 clusters for the NLI dataset. The minimum BIC value was 9 clusters for permutated data. Additionally the BIC curve for NLI data overlapped with the lower boundary for the confidence interval for the permutated data.

This finding suggested that there was no cluster structure to the NLI dataset, consistent with the SLI data and inconsistent with overall Language Impaired data. The confidence interval increased for the NLI data as the cluster size was increased, suggesting that the amount of uncertainty increased, until the cluster size reached approximately 12 clusters at which point the confidence interval envelope began to shrink. Therefore the NLI cluster solutions were not interpreted because the results (1) had a large cluster size and (2) included the permutated minimum in the range of possible solutions.

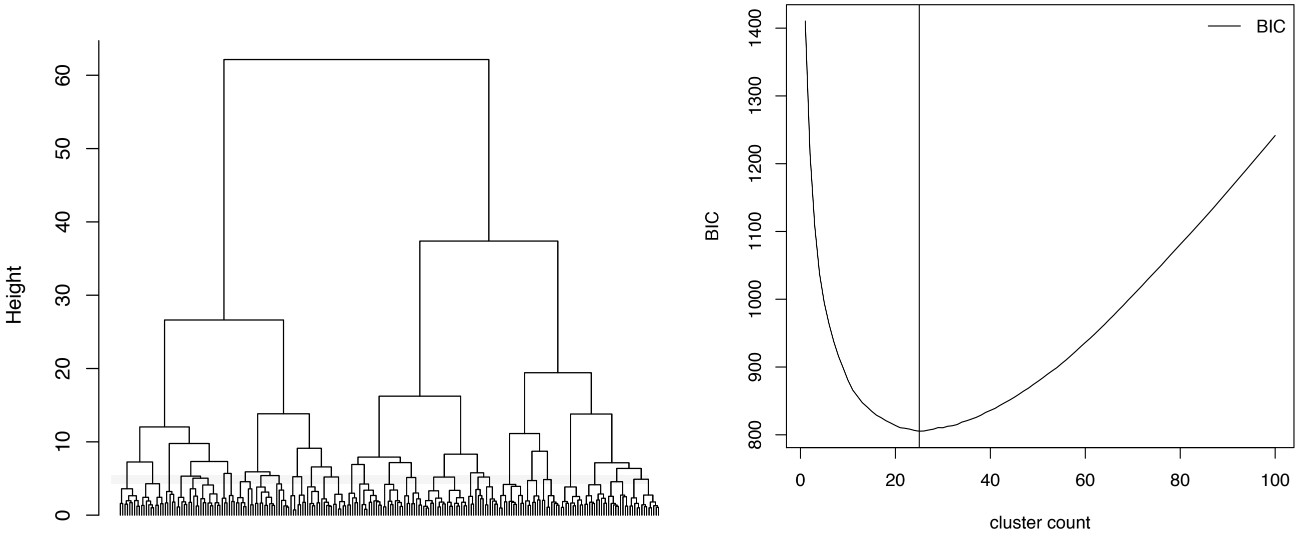


Figure 22 and 23. Ward's hierarchical clustering dendogram (on left) and BIC curve for clusters 1 through 100 (on right) for NLI dataset. Height is the criterion value for a merge and represents the total distance accounted for by that merge. The gray box on the dendrogram and the vertical line represent the identified minimum BIC value.

Principal component analysis was used to reduce redundancy in the variables. To capture 90 percent of the variance at least eight components were used. Figure 25 shows the scree plot of the components. The first component accounted for 22 percent of the variance. The K-means clustering algorithm clustered the data with components one through eight as the variables. The minimum BIC values were 12 clusters for NLI data and 11 clusters for permutated data, larger than the previous clustering results. Figure 26 shows the BIC curves for the principal component results. In contrast to the previous K-means, the BIC curve for the NLI original data was inside the confidence interval for the permutated data. The confidence interval envelope increased as cluster size increased for permutated BIC curve until approximately a cluster size of 30 when the confidence interval began to decrease. The PCA replicated the prior K-means clustering results consistent with the NLI data.

The scatterplot for Ward's method shows a one-cluster solution for of ease interpretation. In contrast the K-means scatterplot presents an eight-cluster solution, which was the mode for the K-means results. See Figures 27 and 28 for the scatterplots. The scatterplots showed no distinct clustering in the data, similar to the BIC results and the Language Impaired scatterplots.

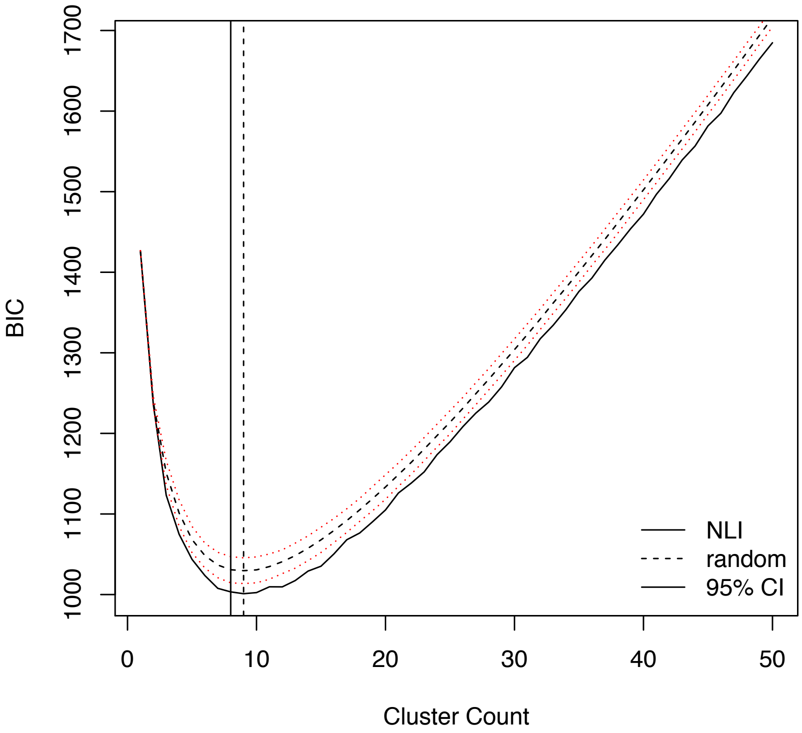


Figure 24. BIC curves for K-means clustering for NLI dataset. BIC curves for original data, permutated data, and 95% confidence interval envelope for permutated curve.

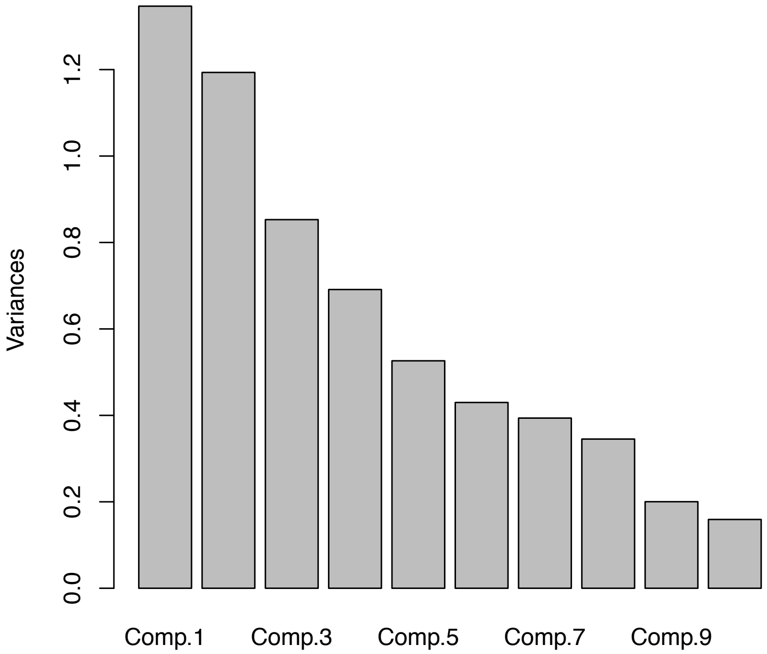


Figure 25. Scree plot of components from principal component analysis for NLI dataset.

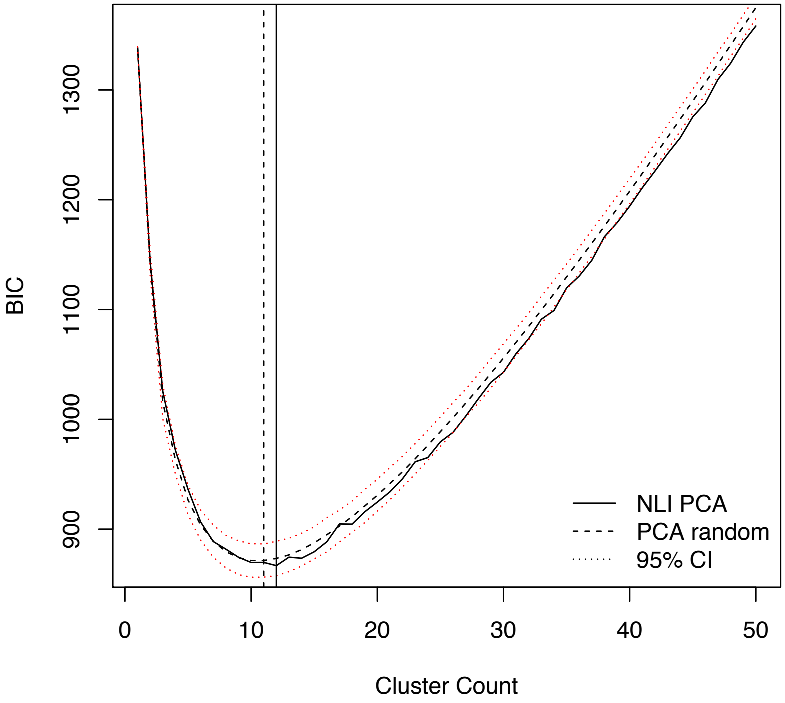
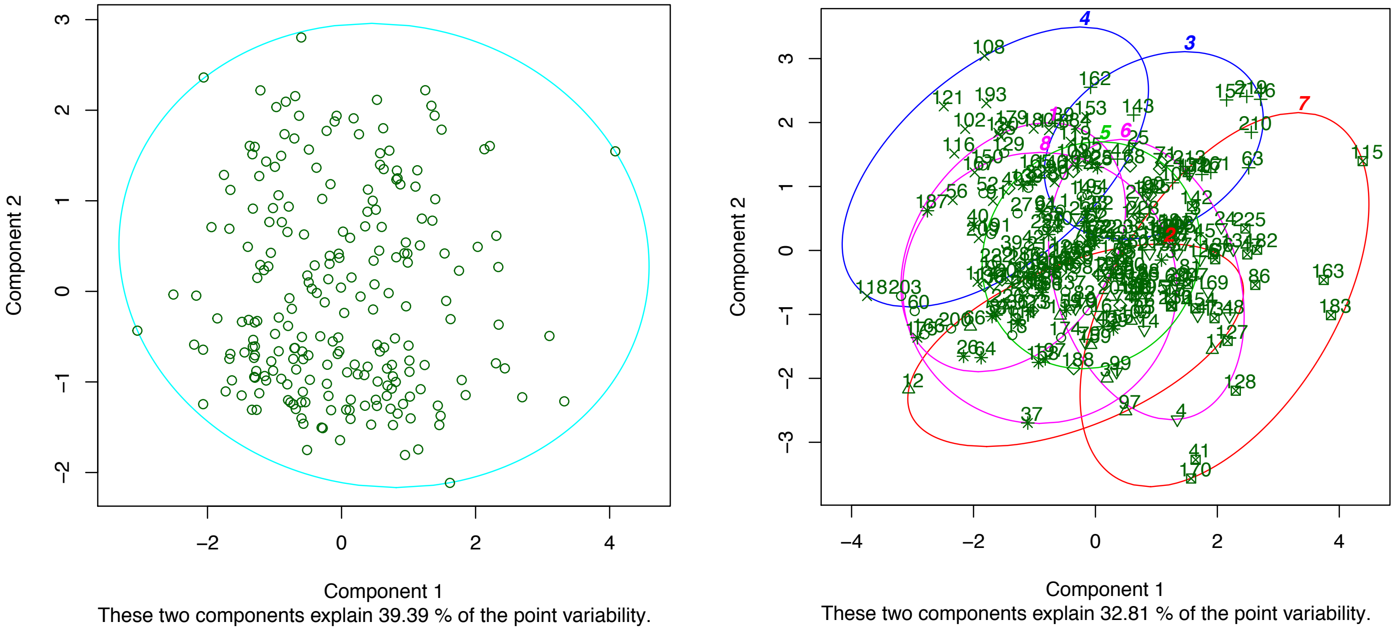


Figure 26. BIC curves when using components for NLI dataset. Curves for PCA results, permutated data, and 95% confidence interval envelope for permutated curve. The vertical lines represent the minimum BIC values.



Figures 27 and 28. Ward's cluster scatterplot (on left) and K-means cluster scatterplot (on right) for NLI dataset.

Profile Analysis

SLI only dataset. Profile analysis was conducted by exploring the means and standard deviations for each variable for the K-mean results. The number of clusters was 10 for the SLI dataset. Regardless that the K-means results appeared to reflect chance, the possible factors were examined to see if there were distinct clusters. See Table 7 for the descriptive information by cluster. The data were graphed in order to visually examine any obvious patterns. See Figure 29 for a line graph of the means by variable for each cluster. Some of the clusters deviated from the other clusters on specific variables. For example, cluster one had the lowest mean for Narrative retell. This separation was a common pattern but did not occur in a distinct pattern, nor did any of the clusters exhibit any clear pattern. This finding was consistent with the K-means results.

NLI only dataset. The possible profiles were examined for the NLI dataset K-means results. There were eight clusters for the K-means results. See Table 8 for the descriptive information by cluster. The data were graphed in order to visual exam any obvious patterns. See Figure 30 for a line graph of the means by variable for each cluster. Although some of the clusters separated on a variable there was no clear pattern for performance. One common pattern, however, that was demonstrated in many of the clusters for several variable pairs, was that a cluster did poorly one of the variables, but rarely both. For example, cluster 3 did poorly on Picture Completion, but comparatively well on Block Design. In contrast cluster 7 did poorly on Block Design and much better on Picture Completion. This finding was consistent with the K-means results for the NLI dataset.

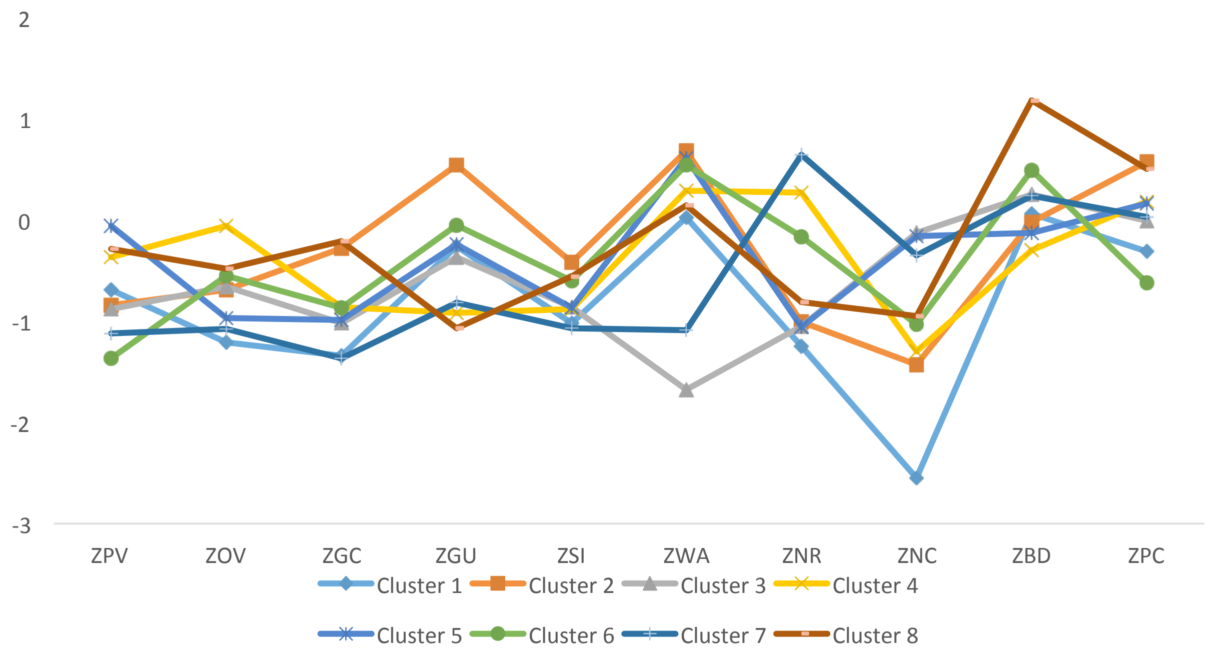


Figure 29. Mean z-score plotted by variables for each cluster for SLI dataset.

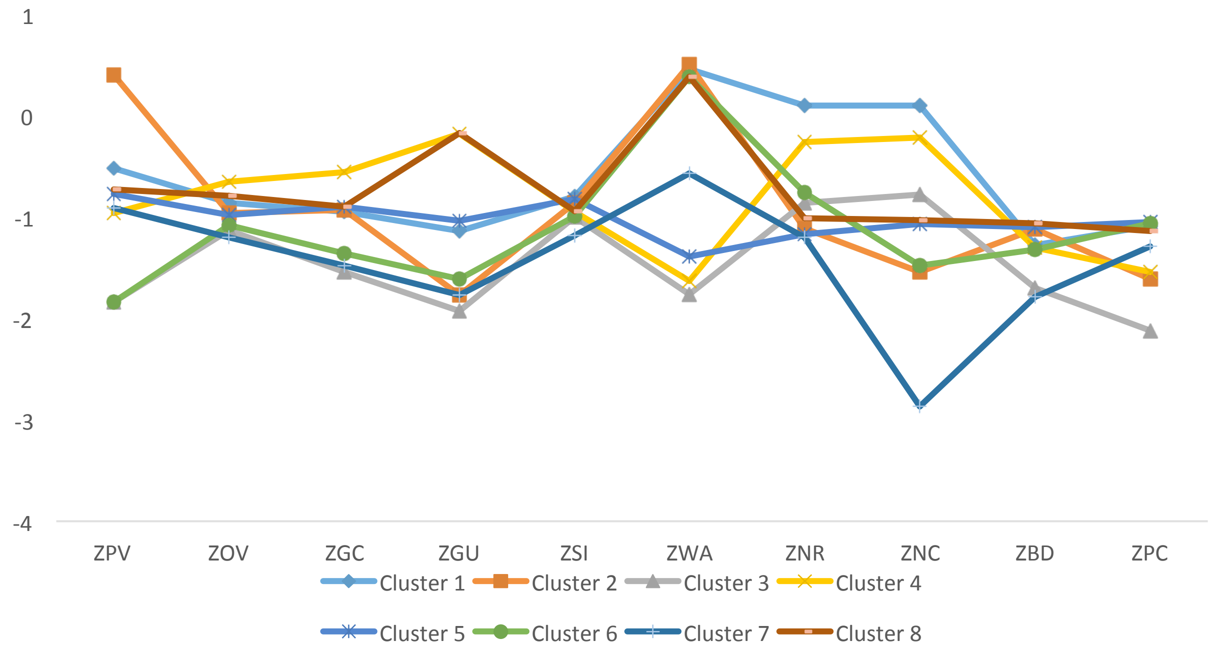


Figure 30. Mean z-scores plotted by variable for each cluster for NLI dataset.

Forced cluster analysis for Bishop's model. We explored the possibility that clinical phenotypes were present in the database and would cluster together in accord with the Bishop classifications. We used the clinical phenotypes that Bishop explored, specifically Typical SLI, Developmental Dyspraxia, Severe Receptive Language disorder, and Pragmatic Language Impairment. The clinical phenotypes were identified in the SLI dataset exclusively for this analysis. The approach for this exploration was to classify cases into Bishop's specified groupings to test if those groupings were concordant with her groupings, and then complete cluster analyses after cases had been classified on the language variables that would aggregate individual children into Bishop's typologies.

Cut-off scores on proxy variables were created to assign cases into the clinical phenotypes from Bishop using the SLI dataset. For Developmental Verbal Dyspraxia (n = 69), the proxy variable was Word Articulation with a cutoff of less than seven standard score.

Table 7

Descriptive information for the clusters for the SLI dataset

	Clusters									
	1	2	3	4	5	6	7	8	9	10
N	27	19	40	26	34	19	19	25	27	41
Picture Vocabulary	-0.69 (0.65)	-0.84 (0.64)	-0.88 (0.62)	-0.37 (0.83)	-0.06 (0.73)	-1.37 (0.46)	-1.12 (0.86)	-0.29 (0.63)	0.16 (0.76)	-1.07 (0.56)
Oral Vocabulary	-1.21 (0.49)	-0.69 (0.60)	-0.66 (0.52)	-0.06 (0.75)	-0.97 (0.51)	-0.55 (0.67)	-1.08 (0.45)	-0.48 (0.58)	-0.82 (0.50)	-1.10 (0.47)
Grammatical Completion	-1.34 (0.45)	-0.28 (0.66)	-1.02 (0.63)	-0.86 (0.64)	-0.99 (0.60)	-0.87 (0.50)	-1.37 (0.49)	-0.21 (0.60)	-0.71 (0.60)	-1.04 (0.56)
Grammatical Understanding	-0.26 (0.81)	0.54 (0.51)	-0.37 (0.71)	-0.92 (0.59)	-0.24 (0.57)	-0.05 (0.57)	-0.82 (0.72)	-1.07 (0.62)	-0.55 (0.84)	-1.69 (0.63)
Sentence Imitation	-1.02 (0.40)	-0.42 (0.74)	-0.86 (0.53)	-0.88 (0.32)	-0.87 (0.36)	-0.60 (0.50)	-1.07 (0.28)	-0.56 (0.64)	-1.09 (0.35)	-0.90 (0.37)
Word Articulation	0.02 (0.72)	0.68 (0.19)	-1.68 (0.50)	0.29 (0.54)	0.61 (0.27)	0.54 (0.27)	-1.09 (0.79)	0.14 (0.72)	-1.59 (0.55)	0.37 (0.54)
Narrative retell	-1.25 (0.49)	-1.01 (0.46)	-1.05 (0.45)	0.27 (0.55)	-1.06 (0.53)	-0.17 (0.63)	0.64 (0.38)	-0.81 (0.54)	-0.93 (0.59)	-0.95 (0.63)
Narrative comprehension	-2.55 (0.65)	-1.43 (0.71)	-0.13 (0.62)	-1.30 (0.46)	-0.16 (0.61)	-1.03 (0.65)	-0.35 (0.64)	-0.95 (0.81)	0.06 (0.68)	-0.52 (0.72)
Block Design	0.06 (0.58)	-0.02 (0.69)	0.25 (0.62)	-0.30 (0.62)	-0.13 (0.55)	0.49 (0.51)	0.24 (0.62)	1.18 (0.64)	-0.12 (0.82)	-0.25 (0.58)
Picture Completion	-0.31 (0.64)	0.57 (0.55)	-0.01 (0.66)	0.18 (0.74)	0.16 (0.58)	-0.62 (0.62)	0.03 (0.78)	0.50 (0.73)	0.84 (0.66)	0.005 (0.71)

Notes. Mean. (sd). N = number of case in a cluster. Values are in z-score format.

Table 8

Descriptive information for clusters for the NLI dataset

	Clusters							
	1	2	3	4	5	6	7	8
N	30	17	15	27	41	35	22	41
Picture Vocabulary	-0.52 (0.80)	0.4 (0.61)	-1.84 (0.53)	-0.96 (0.77)	-0.77 (0.64)	-1.84 (0.61)	-0.91 (0.82)	-0.73 (0.72)
Oral Vocabulary	-0.86 (0.37)	-0.96 (0.52)	-1.13 (0.45)	-0.65 (0.55)	-0.98 (0.42)	-1.08 (0.43)	-1.20 (0.37)	-0.79 (0.57)
Grammatical Completion	-0.95 (0.71)	-0.93 (0.47)	-1.54 (0.43)	-0.56 (0.76)	-0.90 (0.63)	-1.36 (0.43)	-1.48 (0.35)	-0.90 (0.64)
Grammatical Understanding	-1.14 (0.81)	-1.77 (0.69)	-1.93 (0.82)	-0.18 (0.58)	-1.04 (0.58)	-1.61 (0.69)	-1.77 (0.83)	-0.17 (0.46)
Sentence Imitation	-0.80 (0.52)	-0.86 (0.60)	-1.00 (0.71)	-0.95 (0.50)	-0.82 (0.40)	-0.99 (0.38)	-1.18 (0.28)	-0.94 (0.34)
Word Articulation	0.46 (0.47)	0.50 (0.23)	-1.76 (0.59)	-1.63 (0.55)	-1.39 (0.60)	0.38 (0.46)	-0.57 (0.90)	0.38 (0.39)
Narrative Retell	0.10 (0.51)	-1.11 (0.61)	-0.86 (0.73)	-0.26 (0.73)	-1.17 (0.53)	-0.76 (0.59)	-1.20 (0.50)	-1.01 (0.43)
Narrative Comprehension	0.10 (0.55)	-1.54 (0.59)	-0.78 (0.70)	-0.22 (0.71)	-1.07 (0.77)	-1.48 (0.59)	-2.87 (0.49)	-1.03 (0.64)
Block Design	-1.28 (0.52)	-1.12 (0.77)	-1.70 (0.73)	-1.30 (0.65)	-1.10 (0.47)	-1.32 (0.66)	-1.79 (0.57)	-1.06 (0.65)
Picture Completion	-1.09 (0.72)	-1.61 (0.59)	-2.12 (0.69)	-1.54 (0.67)	-1.05 (0.55)	-1.06 (0.67)	-1.29 (0.97)	-1.14 (0.62)

Notes. Mean. (sd). N = number of case in a cluster. Values are in z-score format.

For Severe Receptive Language disorder (n = 10), the proxy variable was the composite receptive score with a cutoff of less than minus two z-score. For Pragmatic Language Impairment (n = 31), the proxy variable was the composite narrative variable with a cutoff of minus
two z-score. For Typical SLI (n = 113), particular participants were classified as Typical SLI if they did not meet the requirements for the other groups, but were otherwise SLI. There were 54 cases not included, because they either could not be classified using this scheme (n = 15) or belonged to two or more groups (n = 39).

SPSS was used to test for between group differences using one-way ANOVAs and post-hoc Tukey tests. Group differences were explored to check if the groups were in accord with Bishop's previous findings. Table 11 includes a summary of the ANOVA results. There were several significant between group differences, however many of the between group differences were not what would have been expected per Bishop's research. For example, there was a between group difference for Picture Vocabulary (F = 8.66, $p < .001$), where the Severe Receptive Language Disorder had the lowest mean (M = 5.5, sd = 1.3), but all the other groups had equivalent means; however, it would be expected that both Severe Receptive Language Disorder and Developmental Verbal Dyspraxia would have differed from the other groups as well as each other. There were two variables were no group differences were found violating expectations. These variables were Oral Vocabulary (F = 0.84 $p = .48$) and Grammatical Completion (F = 1.50 $p = .19$). The finding for Word Articulation most likely represented a skewed finding as all cases for Typical SLI, Severe Receptive Language Disorder, and Pragmatic Language Impairment had to have a value above seven scaled scores.

Ultimately, cluster analysis could not be carried out. There were two main reasons for not performing the cluster analysis. The first was that the overall sample, as well as the smaller clinical phenotype samples, was severely underpowered: A small sample has a higher likelihood of convergence due to chance, which was frequently observed in the SLI and NLI datasets in this study. Furthermore, the small samples of the clinical phenotypes decreased the possibility for separation in cluster analysis. The second reason was that we could not confirm whether or not we had successfully classified cases. This possibility was initially raised because of the cases that belonged to multiple groups; although, overlap is possible in any situation for this analysis it meant that the clinical phenotypes might be too similar when created using these proxy variables. Further doubt was provided by the ANOVA results, as we could not successfully reproduce known differences using the groups created. Therefore, any results obtained would not have been able to appropriately answer the question posed. Moreover, Tomblin and his colleagues did not collect the data to overtly test Bishop's formulation so that our efforts to pre-classify a set of children on measures likely to illuminate groups that fall into Bishops typologies was not successful.

Table 9

Summary of ANOVA results for language variables for "forced cluster" analysis

Variable	**F**
Picture Vocabulary	8.66***
Oral Vocabulary	0.84
Grammatical Understanding	7.46***
Grammatical Completion	1.60
Sentence Imitation	5.33**
Word Articulation	245.20***
Narrative recall	10.81***
Narrative comprehension	30.35***

Notes. * $p < .05$; ** $p < .01$; *** $p < .001$

Supplemental Analysis

The supplemental analyses examined the racial composition of Tomblin's original diagnostic groups to determine if there was a discrepancy from the race proportions for the full dataset. The racial proportions in the full dataset were 84.14 percent White and 13.17 percent Black; these values were used as the expected values for the chi-square tests. Chi-square analyses indicated that the Typically Developing sample ($p = .16$) and SLI sample ($p = .79$) were not different from the proportions for White and Black participants in the full dataset. The chi-square analyses for the Low Normal (i.e., children without a language impairment who also had a nonverbal IQ less than 87 but greater than 70) sample indicated that there was a significant deviation from the expected values ($p < .01$). The Low Normal sample included a higher proportion of Black children (22.22 percent) than the expected value derived from the full dataset. In the NLI group the chi-square analysis indicated a significant deviation from expected value ($p < .001$). Again, the NLI group included a higher proportion of Black children (31.14 percent). Figure 31 provides the proportions for racial groups of the full dataset and Tomblin's diagnostic groups for comparison.

Based on the finding that the low cognitive groups had significantly more Black children than expected, the cognitive scores were examined in more depth. Means and standard deviations were compared to the normative sample for cognitive subtests (i.e., Block Design and Picture Completion) for each racial group (i.e., White/Other & Black). See Table 9 for the means and standard deviations. There was a significant shift for Block Design and a less significant but noticeable shift for Picture Completion for Black. New "race-anchored" z-scores for were obtained for racial groups. Cognitive status was redefined as typical cognition (greater than -0.99) or low cognition (less than -0.99). Using the new cognitive criteria in combination with the original diagnostic category, children were reclassified within language groups (i.e., from NLI to SLI or from Low Normal to Typical). See Table 10 for a comparison of diagnostic group from Tomblin's category to the new categories using the race-anchored cognition scores. Representation was examined in the new samples. Chi-square tests indicated that only the SLI group significantly deviated from the full dataset ($p < .001$) by over representing Black children (23.31 percent) after reclassification.

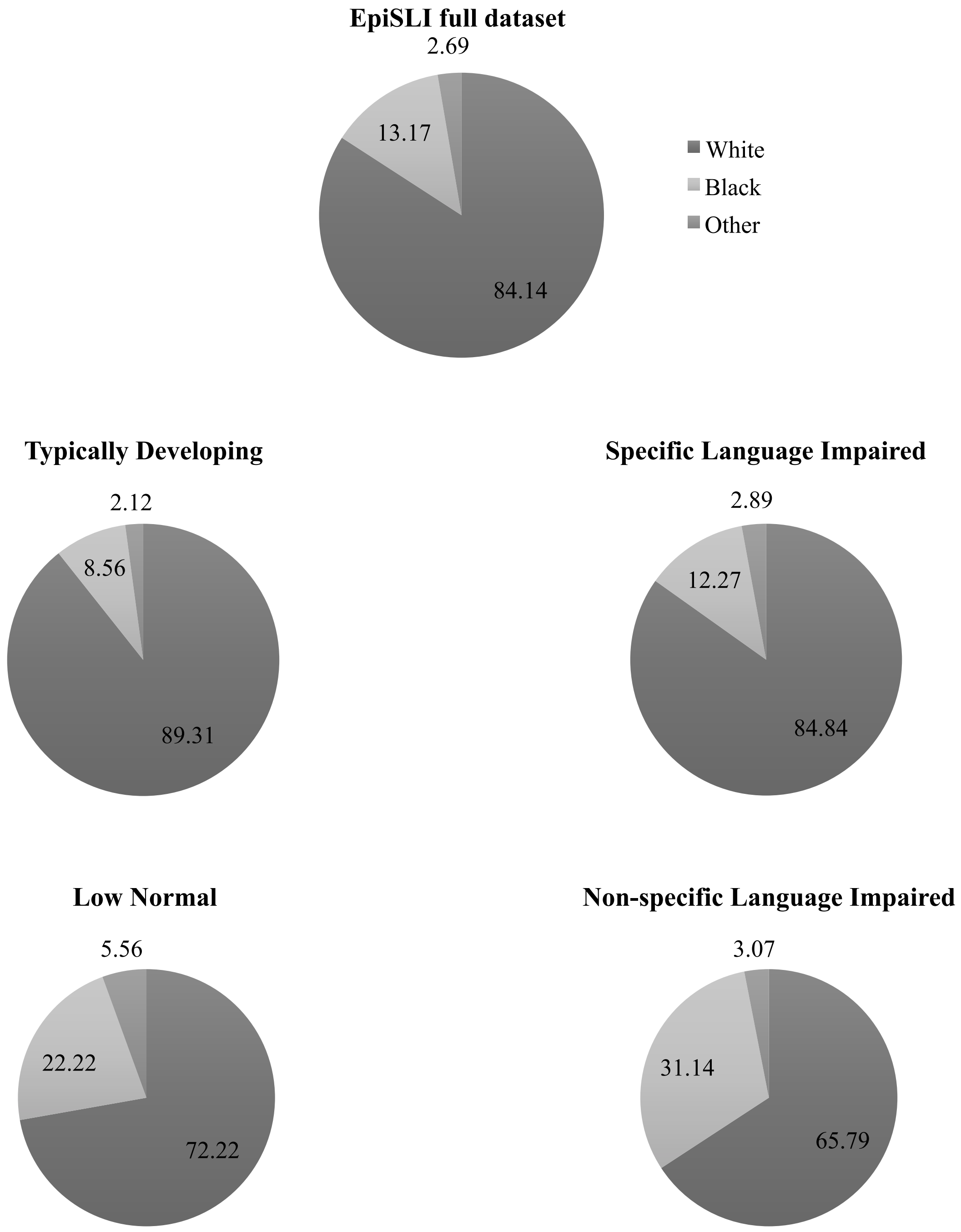


Figure 31. Pie charts depicting race for EpiSLI database and all subsamples within the database. The category "Other" includes Asian American, American Indian, Hispanic, and other minorities.

Table 10

Means and standard deviations by race for cognitive variables

	Variables	
Race	Block Design	Picture Completion
White and Other	9.58 (2.97)	10.34 (2.87)
Black	7.23 (2.93)	8.74 (2.93)

Notes. Means (standard deviations). White and Other were collapsed because of the low sample size for Other.

Table 11

Chi-square table of distribution for diagnostic categories after race-anchored correction

	Diagnostic group after correction			
Diagnostic group Tomblin	Typically Developing	Low Normal	SLI	NLI
Typically Developing	1226	83	0	0
Low Normal	0	115	0	0
SLI	0	0	277	79
NLI	0	0	0	149

CHAPTER 4

DISCUSSION

This study explored subtyping in the EpiSLI database using cluster analysis techniques. Subtypes of Specific Language Impairment (SLI) have inherent potential educational and clinical importance as noted by the Office of Special Education Program (Data Accountability Center, 2012). This study addressed three key limitations the previous studies of subtyping in SLI by including the following methodologies: (a) use of continuous variables without *a priori* number of clusters, (b) adding cognitive variables, and (c) comparing Bishop and Tomblin models which hypothesize subtypes, with the DSM-5 and Leonard SLI models, which hypothesize no subtype solutions. This study had two aims: (1) explore the nature of clustering the EpiSLI database and (2) compare empirical findings to canonical models. Dollaghan (2011) presented the key issue for motivating this study by saying, "the considerable heterogeneity within and between children with SLI has spurred a number of proposals aimed at identifying subtypes of the disorder, but the evidence for meaningful and/or persisting subtypes to data appears weak." (p. 1361). Because my analyses yielded no subtypes, the results of this study support Dollaghan's interpretation of subtyping in SLI.

As noted in the introduction, researchers have been debating whether or not there are subtypes in SLI for many years. The proposed subtypes are thought to be a plausible explanation for a portion of the heterogeneity in children with language impairment. There are different models that have varying numbers of subtypes (e.g., two versus four, Tomblin versus Bishop) and focus on different key features (e.g., pragmatics versus grammar). Despite these theoretical perspectives, there is no consensus on subtype models in SLI and no data driven confirmation of any particular model (Leonard, 2014). Four canonical subtype models were considered in this study. These subtype models were proposed by Tomblin, Bishop, Leonard, and the DSM-5 with the latter two predicting no interpretable subtype solutions.

Tomblin's model had four subtypes of language learners, two of which were language impaired. Tomblin et al. (1997) developed these four subgroups following their epidemiological study of language impairment. The language impaired subgroups differed on nonverbal intelligence. These two subgroups are specific language impaired (SLI) and nonspecific language impaired (NLI). The SLI and NLI groups have been compared in longitudinal studies and these studies found that children with SLI and NLI have differences in many diverse areas of development, suggesting face validity for the Tomblin subtype model. The results of the current study do not support the Tomblin model, which should have yielded at least two distinct clusters in the Language Impaired dataset and sharp SLI and NLI contrast beyond IQ. Instead we found that the solution was 18 clusters for K-means and 51 for Ward's method for the Language Impaired dataset and direct parallels in the SLI and NLI datasets when analyzed independently. These solutions are large and have high overlap in cluster boundaries suggesting that while there may be statistical "clusters," these clusters are non-distinct and inconsistent with the cognitive and linguistic parameters Tomblin posited.

The cluster BIC values are all smaller than chance, even if the minimums were large, which suggests that there is structure, or organization, to the dataset. However, this organization does not support Tomblin's four-subgroup or two-subgroup models and is not seen when the Language Impaired dataset was divided into SLI and NLI only. Interestingly, this study did replicate previous research by Tomblin and Zhang (1999). Tomblin and Zhang clustered the EpiSLI database using the language variables and reported seven non-distinct clusters that had language below normal cut-offs. The current study replicated the finding that there are non-distinct clusters in the data. The SLI and NLI datasets had little organization as is evident from the overlap between the data BIC curve and the lower boundary for the permutated confidence interval. This overlap suggests that the data in those datasets clustered randomly. Tomblin continues to model two separate subgroups, which suggests that Tomblin considers SLI and NLI to be cognitively separate groups but not linguistically separate (Tomblin, 2014).

As discussed in the introduction, Bishop has been exploring subtypes for many years. Her first subtype model was based on clinical impressions. She has translated clinical impressions (Bishop & Rosenbloom, 1987) and incorporated research (e.g., Bishop, 2004) into a refined model. Bishop (2004) subsequently reported evidence for four subtypes of language impairment. I included this model because it provided insight into how clinical impressions inform the understanding of subtypes and Bishop's work has had considerable influence on the field (see Leonard, 2014). However, the findings for this study do not concur with Bishop's four-subtype model. The analysis should have yielded four subgroups in the Language Impaired dataset that were clustered on specific language variables as described by Bishop. As a conservative test of the model, the Language Impaired dataset was reduced to a SLI dataset (i.e., only children with nonverbal IQ composites greater than 87 as suggested by Bishop as a defining trait for SLI). The SLI dataset did not have any clusters, despite being smaller and more homogenous. The EpiSLI database did not confirm Bishop's model because the clusters were most likely due to chance and the minimum corresponded to a large cluster solution. A potential criticism of the analytic approach herein regarding the Bishop finding is the fact that a direct measure of pragmatic ability was not included among the language variables. This caveat has merit for the "pragmatic" subtypes, but even if the pragmatic language impairment subtype was removed, then three clusters should have been evident in the SLI dataset. However, neither a four-cluster nor a three-cluster solution was observed, as the SLI dataset seems to cluster randomly with no specific subtypes. An exploratory forced sample analysis was conducted to further investigate the Bishop subtype model; however the data could not be clustered, because the preparatory phase revealed that the data could not be forced in highly orthogonal groupings.

Leonard, in contrast to Tomblin and Bishop, does not propose that there are subtypes of language impairment. Instead, Leonard argues that children with language impairment had a disruption in the normal language learning process (Leonard, 1989) and can have a wide variety of strengths and weaknesses with a high degree of individual variability. Leonard's model offered a contrasting view that is foundational to argument against subtyping in language impairment (see Dollaghan, 2004, 2011). This study supports Leonard's model because both the Typically Developing and Language Impaired datasets had structure as the BIC curves were better than chance. However, scatterplots showed no distinct subgroupings of the data. Instead there was continuous patterning within each dataset and it might be that there is a continuous pattern in the full EpiSLI database. This result indicates that there is a continuous distribution to the data, which supports Leonard's theory that children with language impairment are the low-end of normal with no distinct (predictable) subtypes. Further support comes from the SLI and NLI datasets, which also had no structure, no unique subtypes and clustered randomly.

The DSM-5 is the current diagnostic for language disorders by the American Psychiatric Association. The DSM-5 also does not suggest subtypes. Rather, the DSM-5 proposed a clinical model that relies heavily on individual strengths and needs. As in the Leonard model, individual differences are key features in the DSM-5 SLI classification. From a theoretical perspective, individual differences would not be expected to yield clinical subtypes; instead one would expect to find either a single cluster or a set of non-distinct clusters. Both of those options would indicate that there were no clinically relevant subtypes. As with the Leonard model and for similar reasons, this study supports the DSM-5; although there was empirical structure in the Language Impaired dataset, the clusters were non-distinct and there were no subtypes.

The main finding for this study is that there are non-distinct clusters and no predictable subtypes in the Typically Developing and Language Impaired datasets. This finding was replicated several times in the analytical approaches applied to the EpiSLI database within this study. Overall, the results support the perspective individual differences, rather than systematic subtypes based on causal mechanisms are the source of the substantial heterogeneity in children with language impairment.

The cluster minimums were greater than twenty for Ward's method in all the datasets. For K-means the clusters obtained fell outside of the 95 percent confidence interval envelope for the permutated data for the Typically Developing and Language Impaired dataset, but within the 95 percent confidence interval envelope for the SLI and NLI datasets. Thus, the BIC curves suggested that there was empirical structure or underlying organization for the Typically Developing and Language Impaired datasets. The underlying organization may be related to continuous relationships between variables instead of between cases. This supposition is supported by the fact that, although consistent reliable clusters did emerge, the minimum clusters were too large to be meaningful. Furthermore, the structure, or organization, did not correspond to any known subtype model in the field of language disorders. Indeed, the organization resembled a normal distribution, which was seen in the scatterplots; further evidence for relationships between variables and not cases. However, the individual SLI and NLI datasets did not have underlying organization, as the BIC curves were not better than chance. The permutated data had more uncertainty for smaller clusters than larger clusters. This uncertainty is evident from the larger confidence interval envelope. The confidence envelope was replicated in all the datasets.

Principal components analysis changed the BIC values. I consistently needed eight components to account for 90 percent of the variance across all the datasets. Across all the datasets the gap between the BIC curves shrank, which indicated that that the clustering became closer to chance. Further, the BIC curve overlapped with the lower end of the confidence interval of the permutated BIC curve for the Language Impaired, SLI, and NLI datasets. Using principal components analysis I created scatterplots of the data used in clustering. These scatterplots included circles that represented clusters, which overlapped or were embedded with other circles for all clustering methods and datasets. Furthermore, none of the scatterplots showed distinct interpretable clusters, which was replicated in all datasets.

The structure, or organization, present in the Typically Developing dataset is demonstrated by the smaller than chance BIC values for all possible cluster sizes, despite the large optimal number of clusters for both Ward's and K-means. This finding is in line with the understanding that, although typically developing children have large individual variance, there are no unique groupings of normal language learners. In the Language Impaired dataset there is underlying organization that is similar to the Typically Developing dataset. The BIC values are smaller than chance, however the cluster sizes corresponding to the minimum BIC values are large indicating that there are no meaningful clusters. In contrast the SLI and NLI datasets had no structure. The SLI and NLI datasets had BIC curves that overlapped or were inside of the lower boundary of the confidence interval for the permutated data. This overlap implies that the clustering results were due to chance. Therefore, the clustering did not yield meaningful clusters. Although cluster sizes were smaller, the clusters were not interpretable.

Throughout the research on subtypes there has been instability and disagreement. The instability and disagreement may be the result of attempting to interpret and impose order on random clusters. This interpretation is plausible because random clustering solutions were systematically replicated multiple times in this study and variable cluster solutions were observed. For example, the NLI dataset had eight clusters, but the BIC curve consistently overlapped with the permutated curve. In fact, when examining the BIC curves for both the SLI and NLI datasets a similar pattern emerged, wherein the BIC curve for the data overlapped or was within the 95 percent confidence interval envelope for the permutated data. This finding indicates that there are no replicable distinguishing features for the SLI and NLI datasets, which appears to be evident in the subtype analysis in SLI (see Leonard, 2014; Dollaghan, 2011).

Supplemental Analysis

This analysis examined racial composition within Tomblin's diagnostic groups. The two groups with "low" cognition included significantly more Black children than would be expected based on the full dataset. Furthermore, analysis of Black children's' scores on the cognitive subtests indicated that the distribution was skewed below the normative distribution. These findings were congruent with previous research (Helms, 1992; Jensen, 2006; Rushton & Jensen, 2006). To correct the skew, children were reclassified using race-anchored z-scores for cognition. The correction positively changed representation for the NLI and Low Normal groups. However, the SLI group had a negative change in representation, changing from representative to discrepant. It is possible that there are children in the language-impaired groups, SLI and NLI, who have normal language but are misidentified due to test bias. This possibility has credence given previous work which has found that Black children are more likely to be identified as language impaired on certain standardized tests (Craig & Washington, 2000).

This finding suggests that within this database there is a bias for Black children on Block Design and Picture Completion resulting in a higher proportion of these children being assigned to the NLI subgroup in the EpiSLI database. The distribution of scores for Black children was skewed from the normative distribution by one standard deviation, but otherwise indicated a bell-shaped distribution. For the EpiSLI database the cognitive tasks were used to identify who was below the "typical" cutoff for cognitive ability. However, in the full database there were two different distributions for the cognitive tasks, White and Black. The different distribution may not be related exclusively to race, but may be related to other socio-economic factors (Cronbach, 1975) or the language demands associated with these tasks (Lancaster & Camarata, 2013). For the EpiSLI database, the racial disparity is important to consider and, when appropriate, be anchored within the proper comparison group (in this case, Black children) in order to reduce misclassification as SLI or NLI. Future studies should use a similar sample relevant scoring system to insure any bias is avoided.

Clinical Implications

The major clinical implication from this study is a confirmation of the importance of determining individual strengths and needs of children with language impairment during assessment. At this time, *a priori* subtyping SLI does not appear to be warranted. The results of this study suggest a focus on heterogeneous individual differences rather than predictable subtypes, in accord with Leonard's model. In light of the results herein it may be useful to examine some of Leonard's assessment recommendations (Leonard, 2014). For example, one such recommendation is the comprehensive examination of all language domains. This project sampled many language domains expressively and receptively, although there were domains that were not sampled as extensively (e.g., phonology) or perhaps at all (e.g., pragmatics). The EpiSLI database provides examples of cases where children have, considered language impaired, highly disparate testing patterns which do not reflect any current subtype models (e.g., Bishop, 2004). Furthermore, this projects supports the assessment of more domains than highlighted by the *DSM-5* (i.e., vocabulary and grammar; APA, 2013). Therefore clinicians rely on diagnostic codes for eligibility and legal compliance, but then focus on a single child's individual profile for determining specific features of the language disorder (Paul, 2007).

Limitations

There are two main types of limitations for the study. The first set of limitations is associated with the analysis procedures. The second set of limitations is associated with the database.

Limitations of analysis. This study employed several statistical forms of clustering analysis. As with any set of statistical procedures there are limitations based on approach. Some of these limitations are specific to the type of clustering method. For Ward's method, one of the limitations can be the order the variables were inputted. In hierarchical clustering the order of the variables can potentially have a large impact on the cluster membership. For example, the clustering algorithm could have been started with the cognitive variables instead of the language variables. This change in ordering may result in clusters more dependent on cognition. However, there is no evidence from this study that changing the order of variables would have significantly impacted the results.

The study accounted for two of the major limitations for K- means clustering, local optimum and initialization points by using permutated data. One limitation of K-means is spherical bias, which was not accounted for in this study. This bias tends to result in clusters that have similar numbers of cases. If this bias did occur it might explain the embedded clusters seen in the scatterplots. However, checking for this bias was beyond the scope of this study. Determining whether or not there was a spherical bias when clustering the data is one avenue for future research.

Cluster analysis using Wards and K-Means has some limitations that apply to all methods of clustering. One limitation is that using different variables always yields different clusters. This limitation can make it difficult to compare across studies on the same population (e.g., language impairment) but with different variables.

Limitations of database. The EpiSLI database provides a unique opportunity to explore subtypes of language impairment there, however there are limitations to the database. A key limitation is that some language variables that are not included (e.g., a direct measure of pragmatics). Some of the language domains needed to make finer distinctions between subtypes, and subtypes models, are not represented in the EpiSLI database. Such variables include more in-depth measures of phonology and pragmatics, as well as additional measures for morphology and syntax. For example, non-word repetition would help to identify which children have difficulties with forming and manipulating phonological information, a key feature for development verbal dyspraxia. Furthermore, Sentence Imitation and Word Articulation do not capture the full breadth of phonology as defined by Leonard (2014). And future studies should include pragmatic assessments that have been developed subsequent to collecting the EpiSLI database. There are a number of different pragmatic assessments that could be used, such as the *Test of Pragmatic Language* (Phelps-Terasaki & Phelps-Gunn, 2007) or a checklist (e.g., Clinical Evaluations of Language Fundamentals Fifth edition; Semel, Wiig, & Secord, 2013).

Another limitation related to measurement is that the cognitive assessment was limited to the subtests on the WPPSI-R. In addition to including verbal instruction and, to a limited extent verbal responses, the optimal database would have included multiple, nonverbal estimates of g. For example, the Leiter-3 includes nonverbal measures of visual-spatial ability and fluid reasoning, as well as a brief IQ based on no verbal instructions.

A separate limitation is that the EpiSLI database is from cross-sectional study. The data represented only one point in time, which makes it difficult to know how the results from this study relate to other important skills, such as early reading comprehension or other academic tasks. Furthermore, the results herein cannot be related to stability of clusters over time. Although it is unlikely that the cluster results from this study are stable, based on the findings here and other research (Conti-Ramsden & Botting, 1999), there was no test of cluster stability over time. The lack of stability testing was a direct result of the nature of the database. To investigate stability a database that had measures repeated over time would be need.

Future Directions

There are several potential future directions for this research line. One direction would be to explore new variables and their effects on clustering. It is possible that results of the current study indicating there were no clusters were attributable to the particular variables and measures used. The domains sampled were reasonable, and likely there was proper domain sampling in the EpiSLI database, but a more refined variable set could be explored. It is clear that the subtype models should have yielded clusters in the current study, even given the limits in measurement within the EpiSLI sample, but including additional variables in key language areas (such as phonology and morphology as well as pragmatics) could potentially increase the likelihood of detecting subtypes.

A second future direction would be to further examine whether or not symptom severity is a dimension on which children with language impairment vary. Leonard (2014) proposed that symptom severity was the only dimension on which children with specific language impairment truly differed on. His proposal is supported the most by the results of this study; however, this study was no designed to investigate symptom severity as a variable. A future study would be to examine how symptom severity might be a grouping characteristic. For example, a measure of grammaticality could be use to symptom severity and be the independent variable for predicting some academic outcome.

Another future direction would be additional large databases. The EpiSLI database is two decades old and was not specifically designed to detect multiple subtypes (such as posited in the Bishop model). Creating a new database designed with the intention of testing for clusters would allow for the best measures to be selected. A new database should be include longitudinal, so as to link early status with long term outcomes more easily, as well as investigate measurement and finding stability.

APPENDIX A.

R SYNTAX WITH EXPLANATIONS

```
# This R script is for Hope S. Lancaster’s dissertation: "Language
disorder typologies: Cluster and principal component analysis in the
EpiSLI database"
# The following packages were used:
# Stats
# Cluster, used for the K-means clustering and scatter plots
library(cluster)
# MatrixStats, used for computing the 95 percent confidence envelope for
the random permutations
library(matrixStats)
# Import data from the EpiSLI database and attach
data.o<- (data.txt)
attach(data.o)

# Remove unneeded variables (e.g., participant id) by deleting columns
data<-data.o[c()] # insert column numbers in the parentheses
# WARD’S METHOD
# For this method use a distance matrix containing squared Euclidean
distances
```

```
# To create the distance matrix for Ward's analysis use “dist” function
and the “Euclidean” method

data.d<-dist(data, method="euclidean")

# Use the “hclust” function with “ward” method for clustering

data.w<-hclust(data.d, method="ward")

# Plot the dendrogram for Ward’s method

plot(data.w, hang=-1, labels=FALSE)

# Calculating BIC for Ward’s method
# To do so use the following functions and run through a loop
# First calculate within cluster sum of squares (wss) function
# This function first groups the cases by cluster
# calculates the mean for variables by cluster
# subtracts individual cases from their cluster mean
# squares the differences
# sums the squares
```

```
wss <- function(clust, data) {
    sapply(unique(clust), function(cl) {
      cl_data <- data[clust == cl,]
      cl_mean <- colMeans(cl_data)
      sum(t(t(cl_data) - cl_mean)^2)
    })
  }

# Second the BIC function for Ward's
# This function requires the wss (ss), p (number of clusters), and n
(number of cases)

bic <- function(ss, p, n)
  2*p*log(n) + ss

# BIC loop
# This loop uses wss and BIC functions to compute the BIC value for 1
through k
# The output of the loop is stored in BIC
# The value for k must be provided
# The value for n must be provided
# This loop identifies cluster membership for 1 through k clusters and
calculates wss for 1 through k clusters
# Then BIC is calculated using wss, number of clusters, and n
```

```
BIC<-NULL
for (i in 1:k)
{
  clust<-cutree(data.w, i)
  data<-data
  out<-sum(wss (clust, data))
  BIC<-c(BIC, bic(out,i,n))
}

# Identify the minimum BIC

which.min(BIC)

# Plot Ward's BIC

plot(1:100, BIC,
     xlab="cluster count", ylab="BIC", type="l")
abline(v=which.min(BIC),lty=1)

# K-MEANS CLUSTERING
# Input data as a matrix for the K-means clustering
data<-as.matrix(data.o[c()]) # remove unneeded variables
# Apply the BIC function for K-means
```

```
}

# The BIC function uses n per cluster (length(km$size)), k
(ncol(sum(km$size))), n for dataset (sum(length(km$size))), and wss
(km$tot.withinss)

BIC.kmeans <- function(km)
  0.5*length(km$size)*ncol(km$centers)*log(sum(km$size)) + km$tot.withinss

# Cluster the data
# The following function clusters the data, calculates, and stores the BIC
values
# data.wss.BIC is the stored variable for BIC values for the data

data.wss.BIC <- sapply(1:50, function(k) {
  BIC.kmeans(kmeans(data, centers=k, nstart=10))
})

# Permutated BIC values
# This function creates the permutated data
k.rand.BIC <- function(x, centers.max=50, ...){
  km.rand <- apply(LI, 2, sample)
  sapply(1:centers.max, function(k) {
    BIC.kmeans(kmeans(km.rand, centers=k, ...))
  })
}
```

```r
# This function clusters the permutated data 250 times per cluster,
calculates, and stores BIC for each run
# iter.max can be increased if needed
# rand.wss.BIC is the stored variable for BIC values for the permutated
data

rand.wss.BIC <- replicate(250, k.rand.BIC(data, centers.max=50,
iter.max=15,nstart=10))

# Identify the minimum BIC values
which.min(data.wss.BIC)
which.min(rowMeans(rand.wss.BIC))

# First calculate 95 percent confidence intervals for permutated BIC
values

mean<-rowMeans(rand.wss.BIC)
sd<-rowSds(rand.wss.BIC)
upper<-mean+1.96*sd
lower<-mean-1.96*sd

# Plot K-means BIC curves
# Placement of legend may need to change depending on curve
```

```
plot(1:50, data.wss.BIC, type="l", lty=1,
     ylab="BIC", xlab="Cluster Count",
     main="Bayesian Information Criterion")
lines(1:50, rowMeans(rand.wss.BIC), type="l", lty=2)
lines (1:50, upper, type=”l”, lty=3, col=”red”)
lines (1:50, lower, type=”l”, lty=3, col=”red”)
legend("topright", c("data", "random", “95% CI”), lty=c(1,2,3), bty="n")
abline(v=which.min(data.wss.BIC),lty=1)
abline(v=which.min(rowMeans(rand.wss.BIC)), lty=2)

# Principal component analysis
# Start with the K-means matrix

data<-as.matrix(data.o[c()])

# Compute components

data.pca<-princomp(data)

# Look at sd, proportion of variance, and cumulative variance per
component

summary(data.pca)
```

```
# Retrieve rotational values for each variable by component

data.pca$loadings

# Scree plot
# Plot amount of variance for each component

plot(data.pca)

# Repeat K-means clustering replacing data with data.pca.k:

data.pca.k<-data.pca$scores

# Select number of components

data.pca.k[,1:4] # replace numbers with the components you want to use

# Scatterplots
# These scatterplots display the data used for a clustering method by 2
components

# Ward's scatterplot uses the distance data

data.w2<-cutree(data.w,k) # replace k with numeric value
```

```
clusplot(data.d, data.w2, diss=TRUE, shade=FALSE, lines=0)

# K-means scatterplot uses the matrix data

data.k2<-kmeans(data,k,iter.max=10,nstart=10) # replace k with numeric value

clusplot(data, data.k2$cluster, color=TRUE, shade=FALSE, labels=2, lines=0)
```

APPENDIX B.

CLUSTERING TUTORIAL

This tutorial provides background for the analyses presented in this dissertation. Examples of the clustering methods are provided using the Fischer iris data.

The Fischer iris data contains three types of irises with 50 cases for each type. The Fischer iris data has four variables. Table 1 shows the mean and standard deviations of petal and sepal dimensions by species.

Table 1

Petal and Sepal dimensions by species for Fischer iris data

	Species		
Variable	Setosa	Versicolor	Virginica
Sepal length	5.01 (0.35)	5.94 (0.52)	6.59 (0.64)
Sepal width	3.43 (0.38)	2.77 (0.31)	2.97 (0.32)
Petal length	1.46 (0.17)	4.26 (0.47)	5.55 (0.55)
Petal width	0.25 (0.11)	1.33 (0.19)	2.03 (0.27)

Notes. Mean (standard deviations).

Cluster analysis is a set of statistical procedures that focus on identifying interrelationships between cases. The primary purpose of cluster analysis is to form clusters, or groups, based on the underlying structure of the data by maximizing the similarity of cases within a cluster and maximizing the dissimilarity of cases between clusters. This tutorial provides background for two clustering methods: **Ward's** and **K-means**.

There are several methods for identifying the optimal clustering solution; this project used **Bayesian Information Criterion** (BIC) because the penalty is larger for BIC than other methods (Schwarz, 1978). **BIC** is based on the likelihood function but includes a penalty for the number of parameters in a model, such as the number of clusters (Schwarz, 1978). The BIC function considers the number of clusters, cases, and the within cluster sum of squares.

Ward's method is one type of **hierarchical** clustering and assumes that each object is unique, but that all of the objects belong to a single overarching category. **Ward's** method uses the squared Euclidian distance information between cases to merge objects together. The merge criterion is the smallest distance between two cases or, at later merges, two clusters. Figure 1 shows the two methods to visually display the results of Ward's method: a dendrogram and the associated **BIC** curve. The dendrogram shows the merges and the criteria value, which is the distance between cases and/or clusters. On the y-axis is the value of the criterion called *height*. Shorter lines represent merges that had low criterion values, whereas longer lines represent merges with higher criterion values. The dendrogram in Figure 1 shows merging with clear possible clusters, as represented by several small merges at the bottom of the graph followed by few large merges from the dashed red line up. The height of the first merge, at the bottom of the dendrogram, was zero, whereas the height of the last merge, at the top of the dendrogram was approximately 200. The blue line represents the height for three clusters, which is the number of species in the data set, and thus the expected number of clusters. The red dashed line is the height identified as the minimum BIC value. The BIC curve is a way to compare clusters to each other to determine which cluster represents the best solution given the model parameters and number of cases. The x-axis of the BIC curve shows the possible cluster sizes, and the y-axis provides the value associated with each cluster size. The cluster corresponding to the minimum BIC values was five, despite there being three species, identified in Figure 1 with a vertical line.

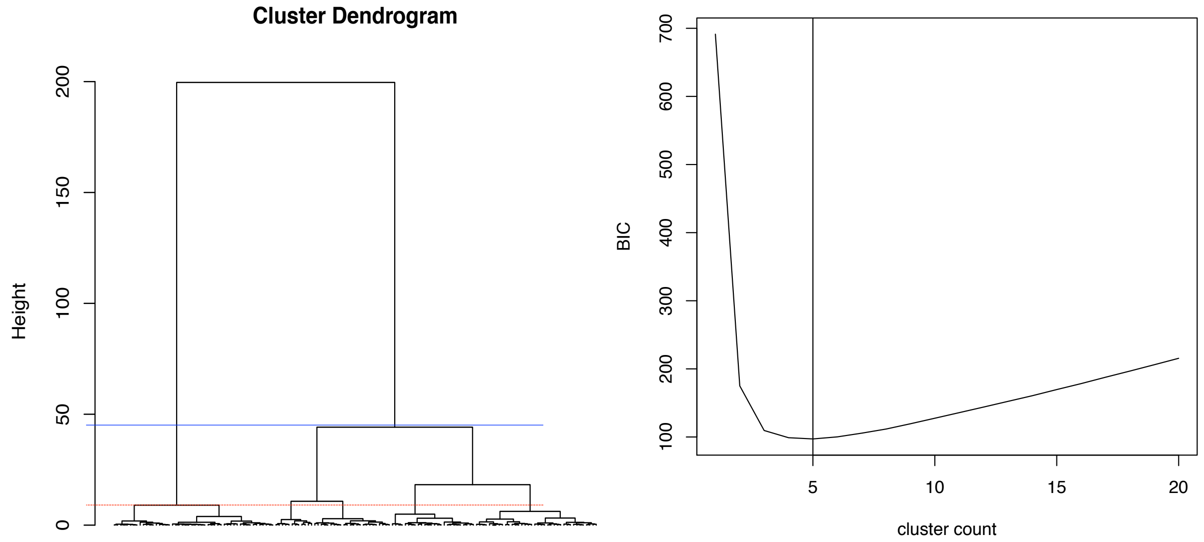


Figure 1. Ward's dendrogram (left) with two different solutions marked and the BIC curve (right) with minimum identified.

K-means is a method of **partitional** clustering and assumes that each cluster is unique and that there is no overarching category. **K-means** method creates centroids, or prototype cases, and then attempts to maximize the similarity within a cluster, while also maximizing the distance between clusters by continuing to reassign until there is model convergence. In order to determine if cluster results were random permutated data was used. The *permutation data* were created using the sample and then clustered using K-means 250 times. BIC values were calculated for each run. The permutated data BIC values were averaged to create a BIC curve and a 95 percent confidence interval was calculated. Figure 2 is the **BIC** curves for the data, permutated data, and the confidence interval for the permutated data. The cluster size corresponding with minimum BIC value is five for the data using K-means and nine for the permutated data. Furthermore, the BIC values are consistently smaller for than chance for the data and do not overlap with the confidence interval. This finding is consistent with the **Ward's method** and indicates that there is structure in the data. However, we know that there are three species in the dataset, which raises the question, what does the structure look like?

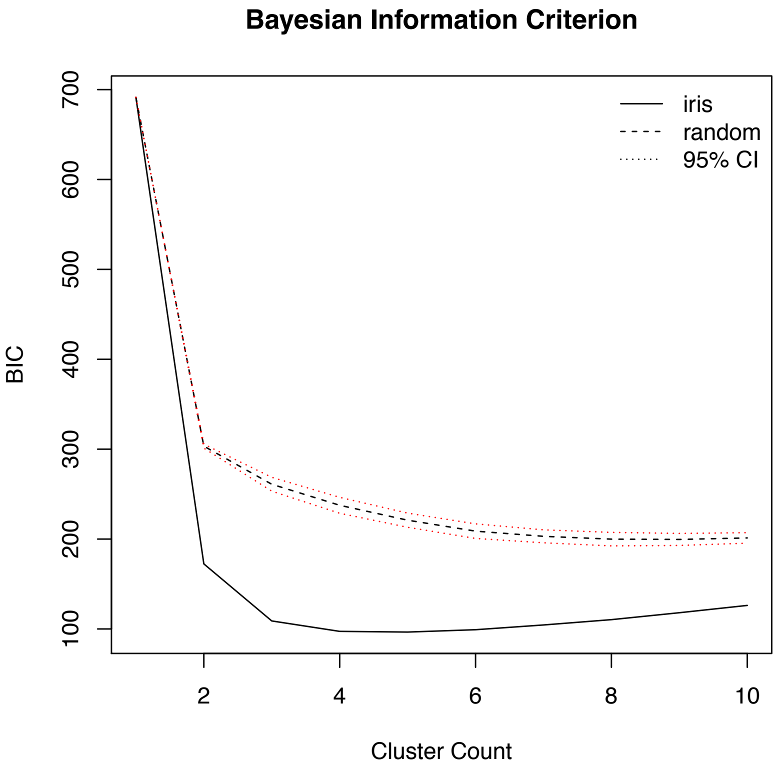


Figure 2. Bayesian Information Criterion values for K-means clustering for data, permutated data, and 95 percent confidence interval.

The structure in the dataset is displayed in the scatterplots. The scatterplots demonstrate why the **BIC** identified five clusters instead of three as the optimum. Figure 3 shows scatterplots of the data used for Ward's (left) and K-means (right). The scatterplots use principal components to plot each case. The circles represent a cluster that was identified using either **Ward's** or **K-means**. Both clustering methods had some error due to the overlap in two of the species. One cluster is cleanly defined; ideally all three species would have a similar degree of delineation on the scatter plot. Although **BIC** overestimated the number of clusters, the scatterplots provide additional information about the structure of the data.

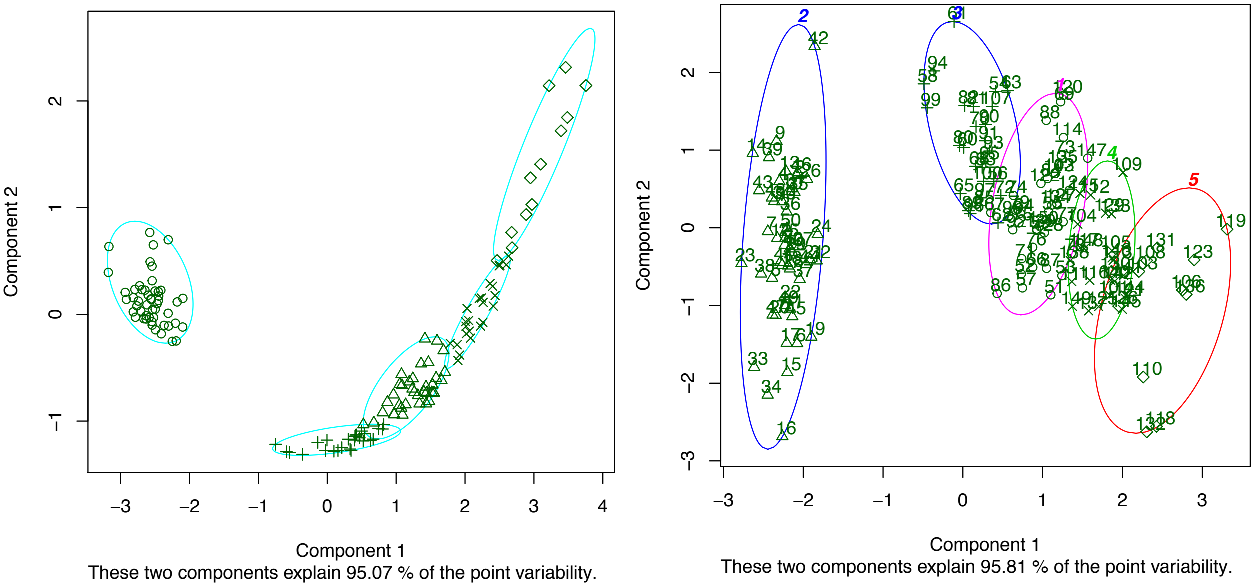


Figure 3. Scatterplots of clusters for Ward’s (left) and K-means (right).

REFERENCES

American Psychiatric Association. (2013). Neurodevelopmental disorders. In *Diagnostic and statistical manual of mental disorders* (5th ed.) (pp. 31 – 86). Arlington, VA: American Psychiatric Publishing.

Bengtsson, H. (2010). matrixStats: Methods that apply to rows and columns of a matrix. R package version 0.2.2. http://CRAN.R-project.org/package=matrixStats

Bishop, D. V. M. (1997a). *Uncommon understanding: Development and disorders in language comprehension in children*. Cambridge, UK: Psychology Press.

Bishop, D. V. M. (1997b). Pre- and perinatal hazards and family background in children with specific language impairment. *Brain and Language, 56,* 1 – 26.

Bishop, D. V. M. (1998). Development of the children's communication checklist (CCC): A method for assessing qualitative aspects of communication impairment in children. *Journal of Child Psychology and Psychiatry, 39*, 869 – 891.

Bishop, D. V. M. (2004). Specific language impairment: diagnostic dilemmas. In L.T. Verhoeven & H. Balkom (Eds.), *Classification of developmental language disorders: Theoretical issues and clinical impairments* (pp. 309 – 326). Mahwah, NJ: Lawrence Erlbaum Association.

Bishop, D. V. M., Bright, P., James, C., Bishop, S. J., & Van der Lely, H. K. J. (2000). Grammatical SLI: A distinct subtype of developmental language impairment. *Applied Psycholinguistics, 21,* 159 – 181.

Bishop, D. V. M., Chan, J., Adams, C., Hartley, J., & Weir, F. (2000). Conversational responses in specific language impairment: Evidence of disproportionate pragmatic difficulties in a subset of children. *Development and Psychopathology, 12,* 177 – 199.

Bishop, D. V. M., & Rosenbloom, L. (1987). Childhood language disorders: Classification and overview. *Language Development and Disorders*, *101*, 16 – 41.

Botting, N. (2005). Non-verbal cognitive development and language impairment. *Journal of Child Psychology and Psychiatry, 46*, 317 – 326.

Botting, N., & Conti-Ramsden, G. (2004). Characteristics of children with specific language impairment. In L. T. Verhoeven & H. Balkom (Eds.), *Classification of developmental language disorders: Theoretical issues and clinical impairments* (pp. 23 - 38). Mahwah, NJ: Lawrence Erlbaum Association.

Catts, H., Fey, M., Tomblin, J. B., & Zhang, X. (2002). A longitudinal investigation of reading outcomes in children with language impairments. *Journal of Speech, Language, and Hearing Research*, *45*, 1142-1157.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad abilities. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jenson* (pp. 5 – 21). Oxford, UK: Elsevier Science, Ltd.

Cole, K., & Fey, M. (1996). Cognitive referencing in language assessment. In K. Cole, P. Dale & D. Thal (Eds.), *Assessment of Communication and Language* (pp. 143-160). Baltimore, MD: Brookes.

Conti-Ramsden, G., & 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., & Botting, N. (1997). The extant to which psychometric tests differentiate subtypes of children with SLI. *Journal of Speech, Language, and Hearing Research, 40,* 765 – 777.

Craig, H. K. & Washington, J. A. (2000). An assessment battery for identifying language impairments in African American children. *Journal of Speech, Language, and Hearing Research, 43,* 366 – 379.

Cramer, D. (2003). *Advanced quantitative data analysis*. Berkshire, England: Open University Press.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychology, 30,* 1 – 14.

Culatta, B., Page, J. L., & Ellis, J. (1983). Story retelling as a communicative performance screening tool. *Language, Speech, and Hearing Services in Schools*, *14*, 66-74.

Data Accountability Center (February, 2012). Office of Special Education Programs *IDEA, Part B* (Revised). Rockville, MD: Westat.

Dollaghan, C. A. (2004). Taxometric analyses of specific language impairment in 3- and 4-year-old children. *Journal of Speech, Language & Hearing Research*, *47*, 464 – 475.

Dollaghan, C. A. (2011). Taxometric analyses of specific language impairment in 6-year-old children. *Journal of Speech, Language & Hearing Research, 54,* 1361 – 1371.

Eddy, K. T., le Grange, D., Crosby, R. D., Hoste, R. H., Doyle, A. C., Smyth, A., & Herzog, D. B. (2010). Diagnostic classification of eating disorders in children and adolescents: How does DSM-IV-TR compare to empirically-derived categories? *Journal of the American Academy of Child and Adolescent Psychiatry, 49*, 277 – 293.

Ellis Weismer, S., Tomblin, J.B., Zhang, X., Buckwalter, P., Chynoweth, J. G. & Jones, M. (2000). Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language, and Hearing Research*, *43*, 865-878.

Feagans, L., & Appelbaum, M. I. (1986). Validation of language subtypes in learning disabled children. *Journal of Educational Psychology*, *78*, 358 – 364.

Heinrichs, R. W., & Awad, A. G. (1993). Neurocognitive subtypes of chronic schizophrenia. *Schizophrenia research*, *9*(1), 49-58.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47,* 1083 – 1101.

IBM Corp. Released 2012. IBM SPSS Statistics for Macintosh, Version 21.0. Armonk, NY: IBM Corp.

Jain, A. K. (2009). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31,* 651 – 666.

Jain, A.K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Systems, 31*, 264 – 323.

Jensen, A. R. (2006) Comments on correlations of IQ with skin color and geographic-demographic variables. *Intelligence, 34,* 128 – 131.

Karasinski, C., & Ellis Weismer, S. (2010). Comprehension of inferences in discourse processing by adolescents with and without language impairment. *Journal of Speech, Language, and Hearing Research*, *53*, 1268-1279.

Lancaster, H. S., & Camarata, S. (2013). *The relationship between pre-literacy skills and fluid reasoning in kindergarteners with and without language impairment*. Unpublished manuscript, Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, TN.

Lambert, E. W., Brannan, A. M., Breda, C., Heflinger, C. A., & Bickman, L. (1998). Common patterns of service use in children's mental health. *Evaluation and Program Planning, 21,* 47 – 57.

Leonard, L. B. (1989). Language learnability and specific language impairment in children. *Applied Psycholinguistics, 10*, 179 – 202.

Leonard, L. B. (2014). *Children with Specific Language Impairment*. Cambridge, MA: MIT Press.

Leonard, L. B. (2009). Is expressive language disorder an accurate diagnostic category. *American Journal of Speech Language Pathology, 18,* 115 – 123.

Leonard, L. B. (2010). Language combinations, subtypes, and severity in the study of bilingual children with specific language impairment. *Applied Psycholinguistics, 31,* 310 – 315.

Leonard, L. B., Deevy, P., Miller, C. A., Rauf, L., Charest, M., Kurtz, R. (2003). Surface forms and grammatical functions: Past tense and passive participle use by children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 46,* 43 – 55.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. (2014). cluster: Cluster Analysis Basics and Extensions. R package version 1.15.3.

Montgomery, J. W. (2002). Understanding the Language Difficulties of Children With Specific Language Impairments: Does Verbal Working Memory Matter?.*American Journal of Speech-Language Pathology*, *11*, 77-91.

Montgomery, J. W., Magimairaj, B. M., & Finney, M. C. (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.

Newcomer, P., & Hammill, D. (1988). *Test of Language Development— Primary 2nd edition*. Austin, TX: Pro-Ed.

Nippold, M. A., Mansfield, T. C., Billow, J. L., & Tomblin, J. B. (2008). Expository discourse in adolescents with language impairments: Examining syntactic development. *American Journal of Speech-Language Pathology*, *17*, 356-366.

Nippold, M. A., Mansfield, T. C., Billow, J. L., & Tomblin, J. B. (2009). Syntactic development in adolescents with a history of language impairments: A follow-up investigation. *American Journal of Speech-Language Pathology*, *18*, 241-251.

Paul, R. (2007). *Language disorders from infancy through adolescence: Assessment & intervention*. Elsevier Health Sciences.

Phelps-Terasaki, D. & Phelps-Gunn, T. (2007). *Test of Pragmatic Language 2nd edition*. Austin, TX: Pro-Ed.

Prior, M., Eisenmajer, R., Leekam, S., Wing, L., Gould, J., Ong, B., & 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*, 893-902.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rasmussen, E. R., Neuman, R. J., Heath, A. C., Levy, F., Hay, D. A., Todd, R. D. (2004). Familial clustering of latent class and DSM-IV defined attention-deficit/hyperactivity disorder (ADHD) subtypes. *Journal of Child Psychology and Psychiatry, 45*, 589 – 598.

Rapin, I., & Allen, D. A. (1983). 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* (pp. 20-35). London: AFASIC.

Rice, M., Tomblin, J. B., Marquis, J., Richman, A., Zhang, X., & Hoffman, L. (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, M., Wexler, K., & 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.

Rushton, J. P., & Jensen, A. R. (2006). The totality of available evidence shows the race IQ gap still remains. *Psychological Science*, *17*, 921-922.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annuals of Statistics, 6,* 461 – 464.

Semel, E., Wiig, E. H., Secord, W. A. (2013). *Clinical Evaluations of Language Fundamentals – 5th edition.* Bloomingto, MN: Pearson.

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology, 59,* 1 – 34.

Stevens, M., Fein, D. A., Dunn, M., Allen, D., Waterhouse, L. H., Feinstein, C., & Rapin, I. (2000). Subgroups of children with autism by cluster analysis: A longitudinal examination. *Journal of American Academy of Child and Adolescent Psychiatry, 39*, 346 – 352.

Swisher, L. & Plante, E. (1993). Nonverbal IQ tests reflect different relations among skills for specifically language-impaired and normal children: Brief report. *Journal of Communication Disorders, 26,* 65 – 77.

Swisher, L., Plante, E., & Lowell, S. (1994). Nonlinguistic deficits of children with language disorders complicate the interpretation of their nonverbal IQ scores. *Language, Speech, and Hearing Services, in Schools, 25,* 235 – 240.

Tallal, P. (1976). Rapid auditory processing in normal and disordered language development. *Journal of Speech, Language, and Hearing Research*, *19*, 561-571.

Tomblin, J. B. (2008). Validating diagnostic standards for specific language impairment using adolescent outcomes. In C. Norbury (Ed.), *Understanding developmental language disorders: From theory to practice* (pp. 93-114). New York, NY: Psychology Press.

Tomblin, J. B. (2010). The EpiSLI database: A publicly available database on speech and language. *Language, Speech, and Hearing Services in Schools, 41,* 108 – 117.

Tomblin, J. B., & Nippold, M. A. (Eds.). (2014). *Understanding individual differences in language development across the school years*. New York, NY: Psychology Press.

Tomblin, J. B., Records, N. L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research, 39*, 1284-1294.

Tomblin, J., Records, N., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, *40*, 1245-1260.

Tomblin, L.B., & Zhang, X. (1999). Language patterns and etiology in children with specific language impairment with specific language impairment. In H. Tager-Flusberg (Ed.), *Neurodevelopmental disorders: contributions to a new framework from the cognitive neurosciences* (pp. 361 – 382). Cambridge, MA: MIT Press.

Tomblin, J. B., Zhang, X., Buckwalter, P., & O'Brien, M. (2003). The stability of primary language disorder: Four years after kindergarten diagnosis. *Journal of Speech, Language, and Hearing Research*. *46*, 1283-1296.

Tomblin, L.B., & Zhang, X. (1999). Language patterns and etiology in children with specific language impairment with specific language impairment. In H. Tager-Flusberg (Ed.), *Neurodevelopmental disorders: contributions to a new framework from the cognitive neurosciences* (pp. 361 – 382). Cambridge, MA: MIT Press.

Whiteman, S. D., & Loken, E. (2006). Comparing analytic techniques to classify dyadic relationships: An example using siblings. *Journal of Marriage and Family*, *68*, 1370-1382.

Watkins, R. V., & Rice, M. L. (1994). *Specific Language Impairments in Children*. Baltimore, MD: Brookes Publishing Co.

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