DEVELOPMENT OF A CHECKLIST
FOR EVALUATING
COHESION IN WRITING

By

Lynda Struthers

B. Sc. Speech-Language Pathology and Audiology,

University of Alberta, 1989

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF EDUCATION

in

CURRICULUM AND INSTRUCTION

University of Northern British Columbia

June, 2001

0-612-80708-8

APPROVAL

Name: Lynda Struthers

Degree: Master of Education

Thesis Title: DEVELOPMENT OF A CHECKLIST FOR EVALUATING COHESION IN WRITING

Examining Committee:

Chair: Dr. Gordon Martel
Assistant (Graduate Studies) to the Vice-President (Academic)
Professor and Chair, History Program
UNBC

Supervisor: Dr. Judith Lapadat
Associate Professor, Education Program
UNBC

Committee Member: Dr. Peter MacMillan
Assistant Professor, Education Program
UNBC

JBell.

Committee Member: Dr. Jim Bell
Learning Skills Centre Co-ordinator
UNBC

External Examiner: Wendy Duke, MSc
Director, Columbia Speech & Language Services, Vancouver
Assistant Professor, School of Audiology and Speech Sciences
University of British Columbia

Date Approved: June 5, 2001

ABSTRACT

This study describes the development and evaluation of a checklist intended for use in the assessment of cohesion in the writing of elementary school children. Assessment of this skill is important as cohesion impacts the readability and quality of written work. Currently available writing tests do not address this area or do so only in a limited fashion. The procedures that I used in evaluating the checklist included classical item analyses, as well as validity and reliability checks. Validity checks provided evidence for construct and discriminant validity. As well, the checklist was able to predict grade membership. Although internal consistency values were low, the level of interrater agreement was satisfactory. Discussion of the findings includes the limitations of this study, suggestions for modifications to the checklist, and future research recommendations.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to thank Judith and Peter who advised me and encouraged me in completing this research. I also wish to thank John, Saima and Willy, my unofficial team of advisors, whose support and knowledge was invaluable. I would also like to acknowledge the speech-language pathologists in School District 57, who participated in many aspects of this study. Finally, but most importantly, I'd like to dedicate this work to my son, Nolan, who shared the first five years of his life with my Masters Degree.

CHAPTER ONE: INTRODUCTION

Assessment may be used for a variety of educational purposes. When preparing for assessment of students, one needs to consider guidelines for best practices. Guidelines outlined in a paper entitled Principles for Fair Student Assessment Practices for Education in Canada (1993) indicate the necessity for test users to select approaches and instruments that are suitable to the purposes of assessment and the students being assessed. Black and Wiliam (1998) echo this sentiment in their discussion of assessment practices in education. Assessment may be used to support a number of educational decisions including selection, placement, and classification of students for acceptance and placement into programs which best suit a student's needs; diagnosis and remediation of particular areas of difficulty experienced by a student; feedback to students; motivation and guidance for learning; and program improvement (Sax, 1997).

As a speech-language pathologist working in an educational setting, I am particularly interested in using assessment for diagnostic purposes. The focus of this type of assessment is on the planning and monitoring of an intervention or instructional program suitable for skill development or remediation. In my professional role I am often concerned with evaluating the writing skills of children with language learning disabilities and difficulties. Children with these types of difficulties often demonstrate problems in communicating their ideas effectively in both spoken and written modes (Wiig & Semel, 1984; Singer, 1995).

In reviewing writing samples of children with language learning problems, it is apparent that many of these children struggle with aspects of writing beyond spelling and grammar. As compared to normally developing peers, these difficulties include: a lower amount of written text produced (Graham, Harris, MacArthur & Schwartz, 1998; Silliman, Jimerson & Wilkinson, 2000), limited diversity in vocabulary, decreased syntactic complexity, less coherence and semantic cohesiveness (Silliman et al., 2000), and poor planning strategies (Graham et al.; Silliman, et al., 2000). I too have noted difficulties with expressing ideas in complex sentence forms and with cohesion in the writing of these children.

Halliday and Hasan (1976) define cohesion as occurring "where the interpretation of some element in the discourse is dependent on that of another" (p. 4). They state that cohesion is a semantic concept, referring to the relationships in meaning that exist within the text. Cohesion is what defines a piece of writing as a unified text and not just a string of words or sentences. The focus of my research is the evaluation of cohesion in writing for diagnostic purposes.

Problem

Assessment Practices

Dagenais and Beadle (1984) reviewed several instruments that assessed writing. Their study included the examination of six achievement tests. These were the Comprehensive Test of Basic Skills, the California Achievement Tests, the Stanford Achievement Tests, the Iowa Test of Basic Skills, The SRA Achievement Series, and the Metropolitan Achievement Tests. These achievement tests focused on the evaluation of word usage, grammar and mechanical aspects of writing. Dagenais and Beadle also examined seven other tests of written language. These included the Test Of Written Language, the Test Of Adolescent Language, Sequential Test of Educational Progress, A Diagnostic System for Teaching Composition for Grades 10-14 (DI-COMP), the Diagnostic Evaluation of Writing Skills, the Test of Everyday Writing Skills, and the Woodcock-Johnson Psychoeducational Test, Part 2 - Achievement for the Written Language Cluster. They found that many of these tests used multiple-choice formats or involved tasks that tested reading more than writing. Dagenais and Beadle indicated that all of these measures involved contrived writing situations rather than naturalistic, authentic writing samples.

A current search of writing assessment tools that I conducted revealed that many tests continue to focus on spelling, punctuation, capitalization, and grammar. Some of these instruments utilized compositional writing samples but still concentrated on scoring mechanical and grammatical aspects of writing. I was interested in finding tests that assessed discourse level structures like cohesion. Discourse level structures are those that reflect the structure and meaning of a text beyond the level of the sentence (Schiffrin, 1994). Those tests that did examine discourse level structures did so in a limited way. These tests provided only one or two ratings for aspects of writing such as organization, sequence, or coherence. A more detailed summary of this search is provided in the next chapter. Dagenais and Beadle (1984) suggested that due to the limited scope of writing skills addressed in commercially available tests, some poor writers may perform adequately on these instruments without actually being able to write more than simple sentence structures or to effectively communicate their ideas in writing.

Commercially available, standardized, norm-referenced achievement tests have their greatest utility in making comparisons between individuals for the purposes of determining eligibility for programs and for predicting success (Sax, 1997). Silliman, Wilkinson, and Hoffman (1993) indicated that traditional approaches to assessment "failed to relate assessment procedures with instructional goals and procedures" (p. 59). They also indicated that these approaches were time consuming yet yielded limited amounts of information for programming purposes. Dagenais and Beadle (1984) indicated that achievement tests are useful in determining who is an "acceptable" writer and who is not. They indicated that none of the tools they examined were intended for in-depth diagnostic work.

In my review of commercially available tests, I noted that many state their purpose as identifying strengths and weaknesses in a student's writing. However, Dagenais and Beadle (1984) state that the practice of using tests for the purpose of simply identifying deficits is both inefficient and unnecessary. Instead, they feel that testing should focus on identifying areas from which to develop teaching programs and compensatory strategies. King-Sears (1994) also criticizes traditional testing indicating that it does not provide information for instructional programming. She states that "norm-referenced, standardized tests provide a snapshot of a student's performance within broad curricular areas, but are not sufficient for developing specific instructional plans when educators must write IEPs [Individual Educational Plans]" (p. 3). Black and Wiliam (1998) also advocate for assessments that provide information for differential treatment of difficulties.

King-Sears (1994) calls for use of assessment materials that analyze errors and provide specific information about where and how to proceed with instruction. Similarly, Rousseau (1990) advocates for the use of error analysis in diagnostic assessments. As children with learning disabilities demonstrate difficulty connecting their ideas in writing (Singer, 1995), that is with cohesion, this may be a useful area in which to focus an in-depth error analysis and diagnostic assessment.

Why Assess Cohesion?

There is a need to evaluate cohesion in writing, as, according to Hedberg and Fink (1996), "errors in cohesion interfere with the reader's efforts to understand the intent of the author" (p. 75). Several researchers have reported that cohesion is related to the overall quality or readability of written work (Lindeberg, 1984; Zarnowski, 1981). Others have indicated a link between writing proficiency and the use of cohesive ties (Englert & Raphael, 1988; Greenberg, 1987; Hedberg & Fink, 1996; Singer; 1995), suggesting that poor writers or those with learning disabilities have difficulty using cohesive ties.

The English Language Arts Integrated Resource Package (IRP) developed by the British Columbia Ministry of Education (1996a) indicates in its prescribed learning outcomes areas that are directly related to cohesion in writing. For instance, the IRPs state that by Grade 4, children will use consistent verb tenses and correct pronoun references in writing and will organize their ideas into logical sequences. These aspects of writing create unity and therefore relate to cohesion. According to the IRP document, by Grade 5 students are expected to be able to revise and edit their own work for clarity. Again, clarity in writing is related to how well ideas, sentences and words are connected for the reader of a written piece and therefore to cohesion.

Nelson (1994) suggests that educators analyze the expectations of the curriculum and abilities of students to develop interventions that narrow the gap between the two. If the goal of writing instruction is to develop writers who can effectively communicate to their readers, cohesion is an important skill and therefore worthy of evaluation. Hedberg and Fink (1996) state that before intervention programs instructing the appropriate use of cohesive devices can be designed, information is required that describes the development of cohesion. Assessment of cohesion in children's writing therefore would not only serve to tell about how a student writes, but also about how cohesion develops.

Summary

For the purposes of my discussion here, current assessment practices are seen to be limited in two distinct ways. First, commercially available standardized assessment measures do not provide a good basis for the development and monitoring of intervention programs. Second, most writing assessments do little to examine discourse level structures that contribute to organization and unity in a written piece. In fact none could be found that addressed specific aspects of cohesion at all.

Aspects of cohesion, on the other hand, are implicated in the curriculum as a learning outcome. It has also been argued that cohesion affects the readability of a written piece. Also, it has been argued that children with language learning problems demonstrate difficulty with cohesion in their writing.

Given these limitations of currently used methods of assessment, and the relationship of cohesion to writing quality, I see a need for an assessment tool that can be used to evaluate cohesion in writing. This instrument should be useful for the development and monitoring of intervention plans and it should be usable with actual writing samples generated by children in the classroom. Development of such a tool would help professionals detect and describe difficulties "problem writers" have in structuring written text (Lindeberg, 1984).

The remainder of this document is devoted to describing the first stages of the development of such an instrument. This chapter provided a brief introduction of the problem to be explored. This problem is elaborated in the next chapter. Chapter Two also provides a literature review into studies of cohesion, assessment of writing, and considerations resulting in the choice of a checklist format for a tool to evaluate cohesion in writing. Chapter Three describes the steps used in developing and evaluating the checklist. The results of these development and evaluation procedures are reported in Chapter Four. This chapter also includes reports of the instrument's reliability and validity. Finally, the interpretations of these outcomes are discussed in Chapter Five. This discussion includes future research directions to continue the development of this cohesion checklist.

CHAPTER TWO: LITERATURE REVIEW

Three key areas are addressed in this review of the literature. The first consists of review of the elements of cohesion and studies of how cohesive devices are used, including in writing done by children. Another area of review focuses on writing evaluation in general. This portion of the review describes how cohesion and writing in general is typically assessed, taking into account both historical and current perspectives. The last area of review focuses on considerations for using a checklist in educational evaluation. Issues of reliability and validity of assessment tools are also examined.

Cohesion

The Concept of Cohesion

The explanation of cohesion provided by Halliday and Hasan (1976) is the most frequently cited in research studies examining the markers of cohesion in children's written and spoken discourse (Crowhurst, 1981, 1987; Pellegrini, Galda & Rubin, 1984; Liles, 1985; Rutter & Raban, 1982; Smith, 1999; Zarnowski, 1981). Halliday and Hasan describe five devices that are used to accomplish cohesion. Examples of each device are presented in Table 1. One device is called reference. This includes the use of pronouns, articles and demonstratives to refer to information within the text (anaphora). Substitution, another device, involves the utilization of a generic term in place of a redundant element. Another tool, ellipsis, involves the elimination of redundant information. A fourth tool, conjunction, is used to connect clauses and sentences and to organize text. Conjunctions may be additive, temporal, causal, adversative, or continuative. A final tool, called lexical cohesion includes lexical reiteration and lexical collocation. Reiteration of a term may be accomplished by using the same word, a superordinate, a synonym or near-synonym. Collocation involves use of words that commonly occur together such as antonyms, complementary terms and converses.

Table 1

Examples of Cohesive Markers

Type of Cohesion		Example
Reference	-pronouns	The boy was cold. He was tired.
	-articles	I saw a dog. The dog started to chase me.
	-demonstratives	A lion stood still. That beast was wild.
Substitution		He always wanted a red bike. Finally he got one.
Ellipsis		I was going to go but (I) didn't (go).
Conjunction	-additive	and, also, in addition, or
	-temporal	then, when, first, next, finally
	-causal	because, therefore, consequently, so
	-adversative	but, although
	-continuative	now
Lexical Reiteration		
	-superordinate	dog - animal
	-synonym	dog - canine
	-near-synonym	dog - beast
Lexical Collocation		
	-antonyms	up - down
	-complementaries	beach - sand
	-converses	ask - answer

Complementary terms are words that commonly occur together. Converses are words that suggest a response of one to the other. The degree of cohesion accomplished through lexical reiteration and collocation is a reflection of the semantic and physical proximity of the terms used in the text. The degree of cohesion is stronger where the distance is less.

The term coherence has also been used by some researchers to discuss aspects of writing related to cohesion. For our purposes here, coherence will refer to the overall semantic unity achieved in a piece of writing whereas cohesion will refer to the linguistic devices used to obtain that unity. For instance, McCutchen and Perfetti (1982) discuss both topic coherence and local connectedness in their discussions of cohesion. According to these authors, topic coherence reflects the overall semantic unity and integrity of a piece and local connectedness refers to the implicit and explicit connections between adjacent sentences. As these authors explain, topic coherence is necessary but does not in and of itself create coherence in writing. They state that it is difficult to describe overall global coherence without describing the devices used to establish connections between sentences. These local connections between sentences reflect the same cohesive devices described by Halliday and Hasan (1976).

Cohesion and writing ability. Cohesion has been shown to be related to the overall readability and quality of written language. Zarnowski (1981) cited the relationship between inter-sentence cohesion and the readability of a written text in her argument about the importance of analyzing cohesion in children's writing. Rutter and Raban (1982) stated, "Failure to realize the implication of a cohesive tie, to recover its referent, implies loss of meaning and a break down in coherence for the recipient of the communication" (p. 65). This relationship between readability and cohesion is further supported by an exploratory study conducted by Lindeberg (1984) where graded college level expository essays were analyzed for the use of cohesive ties. She found that the proportion of cohesive ties was greater in essays graded 8 or more out of 10 than in those that received grades lower than 6. She concluded that her findings supported the hypothesis that cohesive tightness could be considered a sign of quality in writing.

Not only is cohesion related to the overall readability of writing but it has also shown a link to writing proficiency. Singer (1995) reported that children in Grades 3, 5 and 7, without a history of language-learning difficulties, were "remarkably adept" at writing cohesively. Englert and Raphael (1988) indicated that children with language-learning impairments have difficulty detecting inconsistencies in their writing and recognizing how these inconsistencies confuse the reader. They concluded that such difficulties would be expressed as problems in coherent organization of ideas in written prose. In an investigation of cohesion in the writing conducted by Hedberg and Fink (1996), normally developing children were compared with children with language-learning disabilities. Children with language-learning disabilities scored significantly lower than their normal peers on many of the variables examined in their writing. Included in these lower scores were demonstrations of less cohesive harmony and density than their peers. In a study of narrative and expository writing samples from children in Grades 2, 4, 6, and 8, McCutchen and Perfetti (1982) found increasingly higher percentages of inter-sentence cohesive ties used with increased grade.

The relationship between the readability of a written composition and cohesion that is cited in the literature highlights the importance of understanding and evaluating cohesion in writing. This argument is furthered by evidence that writing proficiency is related to the ability to write cohesively.

Cohesive Devices Used by Children in Writing

Although the research literature is not extensive in this area, some researchers have undertaken investigations of the types of cohesive devices used in the writing of children (Crowhurst, 1981, 1987; Hidi & Hildyard, 1983; McCutchen & Perfetti, 1982; Pellegrini et al., 1984; Rutter & Raban, 1982; Smith, 1999). These studies indicate that certain cohesive devices appear more frequently in the writing of children, whereas others appear infrequently or not at all. For instance, Crowhurst, Liles (1985) and Smith found that substitution and ellipsis were rare or absent in the writing of the elementary school-aged children studied. Crowhurst and Smith rarely encountered continuative conjunctions and Liles noted a lack of comparative reference. In a small study conducted previously, I found that pronoun referencing; lexical cohesion; and causal, temporal and additive conjunctions were the most frequently used cohesive devices in the writing of children in Grades 3, 5 and 7. Crowhurst's studies of writing by students in Grades 6, 10 and 12 showed that the most common cohesive ties used were lexical cohesion, pronoun referencing, demonstratives, and use of the definite article. Liles investigated spoken rather than written narratives of children aged 7 years 6 months to 10 years 6 months and found a greater percentage of reference and conjunction than other cohesive devices.

Some interesting findings related to the uses of reference and lexical cohesion. For instance, while pronoun referencing appeared a predominant cohesive device used by children in several studies, in my study I noted errors in referencing pronouns in the writing samples of elementary school-aged children (Smith, 1999). I also noted that pronoun referencing and lexical cohesion were among the most sensitive to developmental variation. That is, these devices were used differently at different grade levels. Rutter and Raban (1982) found differences in the way children of different ages used demonstratives. In their study of narrative writing, 10 year olds used a higher proportion and greater variety of demonstratives than did 6 year olds. They also found that the lexical device of collocation was used more frequently than superordinates for both age groups.

Studies of cohesion have similar findings with respect to the conjunctions that are commonly seen in the writing of children. In a study by Crowhurst (1981) the additive conjunctions and, also, and the adversative conjunction but were the most commonly used. So was the most commonly used causal conjunction (Crowhurst, 1981, 1987). Scott (1991a) reported that inter-sentence connections in the narratives of students are usually accomplished with so and then. Crowhurst (1987) also noted use of temporal markers such as then, soon, later and next day. Similarly, I found the additive conjunction and to be used in the writing samples of all the children in my study, whereas so was the most common causal conjunction. I also found frequent uses of the additive conjunction also, the causal conjunctions if and because, and temporal connectives then, when and before/after in descending order of frequency of appearance. Overall, causal, temporal and additive conjunctions were the most commonly used (Smith, 1999). I also noted that the kinds of additive, causal and temporal conjunctions used change across grades, with older children using a greater variety.

Aspects of cohesion which provide unity across the text have not been as well studied. Perrara (1984) indicated that use of consistent tense across the text provides important discourse connections. She indicated that younger writers quite often have difficulty with this aspect of writing. I found that topic coherence was primarily realized through lexical devices used across the text, and through the organization of the text which was achieved through paragraph structure and sequencing of information (Smith, 1999).

The predominance of certain cohesive devices used in the writing of children suggests that these may be important areas to focus on in an analysis of cohesion. These include pronoun referencing, demonstratives, and use of the definite article the. Conjunctions in the additive, temporal, causal, and adversative categories also may be important to examine. Another area worthy of examination includes aspects of lexical cohesion with a focus on the use of collocation and superordinates. A final area to consider would be aspects of global cohesion such as paragraph structures and other organizational features.

Methods of Evaluating Cohesion in Used Research

Liles (1985) criticizes the simple classification of cohesive devices as done in the research cited thus far as insufficient for analyzing cohesion. She suggests that studies of cohesion should also address the issue of cohesive adequacy. In her study of cohesion in the oral narratives of primary school children, she utilized a process in which cohesive ties were identified, categorized and then judged as complete, incomplete, or erroneous. The raters in her study identified cohesive markers by reading each sentence of the transcript in isolation. An element was considered cohesive if the reader had to search outside the sentence for its interpretation. The classification of ties was then made according to the definitions of Halliday and Hasan (1976). A tie was considered complete if the referred information could be determined unambiguously. An incomplete tie involved interpretation that appeared to be based on information that was not provided in the text. An erroneous tie resulted when the listener or reader was guided to ambiguous or erroneous information. In the evaluation of conjunction use, conjunctions were judged to be either complete or erroneous as it was considered too difficult to judge the completeness of a conjunction.

This procedure for identifying cohesive markers creates difficulty in measuring lexical cohesion. This type of procedure would not be sensitive to the use of complementary terms and converses. These types of markers contribute to cohesion by creating semantic relationships across a piece of writing. However, the interpretation of each element is not dependent on the other. This method would only capture examples of lexical reiteration. It also fails to examine global aspects of cohesion as it does not examine elements like consistent use of tenses or overall organization of ideas.

Another method of measuring cohesive adequacy was employed by Liles, Duffy, Merrit, and Purcell (1995). These researchers measured adequacy by dividing the number of complete ties by the number of ties in each linguistic category. Identification, categorization, and judgment of tie adequacy followed the procedure laid out by Liles (1985). Again this method omits analysis of aspects of lexical and global cohesion.

In a study by Klecan-Aker and Lopez (1985), only reference and conjunction were measured. The authors failed to indicate why other areas of cohesion were not examined in their study. In their analysis, they described reference as either appropriate or inappropriate. Appropriate ties were those that were unambiguous. These authors indicated that one factor which reduces ambiguity is the close proximity of a reference to its referent. Conjunctions were measured by first being categorized as coordinating or subordinating. The number of each type was then counted for each writing sample. In my view, this approach does not seem to be measuring conjunctive cohesion as much as syntax. The classification of conjunctions as coordinating or subordinating reflects the syntactic complexity of the written piece but does not provide information as to the kinds of relationships between ideas, that is whether they are causal or temporal, for example.

A method for analyzing cohesion in writing based on these and other pieces of research was described by Hughes, McGillivray, and Schmidek (1997). As in the methods described above, cohesive ties are identified when they refer to information somewhere else in the text. The ties are then judged and counted as suggested by Liles (1985) and Liles et al. (1995). Hughes et al.'s adaptation to these procedures lies in the method of preparing a sample of writing for analysis. Their procedure involves dividing the writing sample into main clauses with subordinating clauses attached (T-units) rather than dividing it by sentence boundaries. As the remainder of the method does not vary significantly from those already described, the same criticisms as for the methods used by Liles and Liles et al. apply here.

The methods of analyzing cohesion described here were developed primarily for research purposes. Three considerations for evaluating cohesion have been highlighted in these methods. First, in evaluating cohesion, it may be important to consider whether ties are ambiguous or clear. Second, the proximity of ties should be considered in their evaluation. Third, establishing T-unit rather than sentence boundaries may be helpful in evaluating inter-sentence cohesion.

Summary

This review of studies of cohesion highlights the need for a clinical instrument that could be used to evaluate cohesion in writing. Furthermore, these studies suggest content areas that might be included in such an instrument. Studies relating cohesion to the readability and quality of writing highlight the importance of cohesion as a subject for assessment. Research depicting how cohesion is used in the writing of children provides information about what types of devices to assess and provides considerations of methods that may be used in developing an assessment tool for cohesion.

Writing Assessment

Writing assessment procedures vary across instruments and across time. Evaluation of writing has followed trends in assessment practices that reflect changes in the social climate, trends in research, and shifts in educational practices. A few of these influences on writing assessment are discussed here, along with an exploration of current methods of writing assessment.

Historical Perspective

Earlier in the twentieth century, assessment practices centered on standardized objective measures of learning. Several factors contributed to this approach. One was the development of the multiple-choice technology during the World Wars, to allow for inexpensive and efficient selection of soldiers (Calfee & Freedman, 1996). This coincided with an emphasis on accountability and behavioral approaches. Standardized, multiple choice testing meshed well with this purpose (Calfee & Freedman, 1996). In the mid 1950s there was also a move to make assessment in education more objective through the use of indirect measures that emphasized right versus wrong (Isaacson, 1991). The focus of writing evaluation at this point related to objectives-based education with an emphasis on spelling and grammar (Calfee & Freedman, 1996). This trend was facilitated by the development of machine scoring (Isaacson, 1991). Evaluation of writing using this approach, however, proved to be problematic. First, writing was not easily assessed by these methods. In addition, reliability of standardized writing measures was difficult to achieve (Calfee & Freedman, 1996).

Eventually, perspectives about writing and writing evaluation began to change. In the early 1970s, two projects, namely the Bay Area Writing Project and the National Writing Project, placed emphasis on the concept of "writing as a process" (Calfee & Freedman, 1996). In addition to this new perspective, the difficulty in assessing compositional writing skills through indirect measures sparked an interest in holistic scoring (Isaacson, 1991). Thus writing came to be seen as a complex process that required direct, holistic examination.

Despite these changes in the viewpoints around writing and writing assessment, research and practice in areas of language intervention still operated from a behavioral viewpoint throughout the 1970s (Warren & Yoder, 1994). This finally began to change in the

1980s for several reasons. One of these reasons was a movement from the stricter behavioral point of view towards naturalistic contexts for learning (Warren & Yoder, 1994). That is, researchers began to see that skills trained in a strictly behavioral fashion did not generalize well. Another reason for the shift in assessment and intervention practices related to the whole language movement in the 1980s. This movement emphasized a more authentic curriculum in both reading and writing (Calfee & Freedman, 1996). The emphasis in writing with this movement included aspects of writing such as purpose, voice, audience, and coherence (Calfee & Freedman, 1996). These changes to writing instruction were connected to a shift in how writing was assessed. The whole language theory meant that labels were not as important as descriptions, and learning was related to context (Gillam & McFadden, 1994). With the focus on context and the view of writing as a process, authentic assessment, performance-based testing, and the use of portfolios for evaluation began to emerge (Calfee & Freedman, 1996). Coinciding with these changes was a movement to revise holistic scoring procedures to include countable features, thus balancing the need for objectivity with the need to evaluate authentic writing tasks (Isaacson, 1991).

The goal to balance objective measures of writing with the need to observe writing directly for evaluation purposes continues to challenge test developers. This can be seen in examining current methods and tools for assessing writing.

Commercially Available Tests

A search of currently available assessment tools was undertaken to analyze which aspects of writing are presently addressed in tests of writing. As the availability of assessment tools for direct review was limited, test indices were used to provide a fairly exhaustive review of tools that may be available for the assessment of writing skills in elementary aged students. Review of tests was conducted using four main indices. These were Tests in Print V (Murphy, Impara, & Plake, 1999), Psychological Assessment in Schools (Impara & Murphy, 1994), The Thirteenth Mental Measurements Yearbook (Impara & Plake, 1998) and the ETS Test Collection Catalogue Volume I: Achievement Tests and Measurement Devices, 2nd Edition (Educational Testing Service, 1993). A web search was also conducted using a test locator through ERIC. Whenever possible, a review of the actual test was conducted.

The focus of this search was on finding tests that evaluate written language in actual narrative writing samples of elementary school children. These limitations were placed on the search as they reflected the parameters set out for this study. As per the findings of Dagenais and Beadle (1984), many of the tests found that addressed writing at all focused on indirect writing measures (e.g., sentence completion, word and sentence writing, cloze activities) or on mechanical aspects of writing such as punctuation, spelling, and capitalization. Tests that did examine direct writing samples usually utilized holistic or analytic scoring procedures. Holistic scoring involves rating a whole written text with a single score. Analytic scoring involves rating several aspects of a written piece individually. These terms are defined in more detail in the upcoming section on rating systems. Sax (1997) criticizes holistic measures as being too subjective. Murray-Ward (1998) indicated that holistic scoring is useful in a general writing assessment but does not allow for diagnostic information. In her opinion, analytic scoring was more useful in examining aspects of writing in a more isolated fashion and could be useful to assist students in improving particular aspects of their writing.

Even when analytic scoring was used, it was often used to examine mechanical aspects of writing or provided only one or two ratings for discourse level structures such as organization, sequence or coherence. For example, the CTB Writing Assessment System, as described by Engelhard (1998), uses analytic scoring for content, organization, sentence construction, vocabulary/grammar and spelling/capitalization. Other tests that use scoring in this way include the Test of Written Expression (TOWE) and the Test of Written Language (TOWL-3). The TOWE analytically scores attributes such as organization and structure, detail, spelling, punctuation, capitalization, usage (Murray-Ward, 1998). The TOWL-3 analytically scores Contextual Conventions, Contextual Language, and Story Construction (Hansen, 1998). Hansen (1998) criticizes the TOWL-3 for using judgments that are too subjective like "poor", "average" and "good" to rate writing samples.

Only three instruments were found that assessed coherence. These were the Writing Process Test (WPT), the Wechsler Individual Achievement Test (WIAT), and the Oral and Written Language Scales (OWLS).

As described by Kimmel (1998), the WPT utilizes a two-phase procedure. In the first phase, the writer is required to plan and draft a composition. The second phase involves editing and revising the first draft. The draft is then scored with an extensive five-point rating system on various features of writing competence. These include aspects of writing such as purpose, audience, vocabulary, style, and mechanical aspects of writing such as punctuation and spelling. A rating is also provided for Organization/ Coherence. With this rating, coherence and organization refer to how well the writer adhered "to a discernible plan throughout the composition"(p. 1160). This does not appear to be related to the definitions of coherence and cohesion examined earlier in this chapter, however.

The WIAT (Psychological Corporation, 1992) is administered by having a student write about a topic for 15 minutes. The composition is then scored using both holistic and analytic rating scales. One of the analytic ratings scores Organization, Unity, and Coherence on a four-point scale. Criteria for a rating of 4 is described as "Completely organized, with smooth flow from one idea to the next through the use of transitions and sequencing. Unity is strongly evident with no wandering from the primary theme or plan" (Psychological

Corporation, 1992, p. 74). A rating of 1 is given to samples described as "Lack of plan. May be incoherent" (Psychological Corporation, 1992, p. 74). This rating serves as one score of an analytic scoring system that examines six elements of writing. The WIAT also uses a six point holistic measure on the same sample of writing with the top score including criteria for unity and organization of the piece.

The OWLS (Carrow-Woolfolk, 1996) consists of a variety of writing tasks including writing sentences and paragraphs. This test also examines coherence in writing. However, the test contains only two items that test coherence for elementary-aged children. Each of these items receives a score of either 1 or 0 for the presence or absence of coherence in a short writing sample. This test gives credit for coherence when each sentence is tied to the previous one, tenses are consistent, transitions like then, next, and so forth are used, and sentences are not "choppy."

None of these instruments provided an in-depth analysis or definition of each of the linguistic tools used to achieve cohesion in writing. Each test analyzed the more global concept of coherence. Furthermore, in each case, only one or two item scores reflected this aspect of writing competence. While analytic scoring has been viewed as diagnostic in nature, it is my contention that it would be difficult to develop interventions aimed at improving skills in cohesion or coherence based on a single rating of the overall skill.

Methods of Assessment Using Curriculum-Generated Writing

Some methods of assessment involve using real writing samples generated through regular classroom assignments and applying a scoring system to them. When evaluating writing in this way a variety of measurement procedures are possible. Silliman and Wilkinson (1994) discuss several options for the evaluation of language skills by observing their use in regular classroom tasks. For our purposes, this would translate to examination of curriculum-generated writing. Several methods used for this kind of assessment will be examined here including curriculum-based assessment and categorical tools.

Curriculum-Based Assessment

Many methods of evaluating samples of written language are referred to as forms of curriculum-based assessment (CBA) or measurement (CBM). Poteet (1992a) defines CBA as “the process of determining students’ instructional needs within a curriculum by directly assessing specific curriculum skills” (p. 11). CBM is a specific set of procedures for repeated measures of student progress on standardized tasks of writing expression (School District #57, 1996). The term CBA implies an overall approach to assessment, while CBM refers to a particular set of measures. According to Choate and Miller (1992) CBA determines the expectations of the curriculum, the match between students and those expectations, and how to plan to adjust the curriculum to meet the needs of students. Nelson (1994) also describes a similar procedure which she refers to as curriculum-based language assessment. She distinguishes this form of assessment from CBA in that CBA addresses whether or not the child has learned the curriculum whereas curriculum-based language assessment determines whether or not a child has the language skills and strategies necessary for processing the language of the curriculum.

As the models of CBA are varied (Poteet, 1992a; King-Sears, 1994), so are the CBA methods of assessing writing. Despite these variances, some general guidelines are supported by several researchers. For instance, Nelson (1994) indicates the need to use the real context and content of the curriculum in assessment. King-Sears (1994) recommends that quantitative and qualitative measures used in CBA reflect the teaching objectives. Similarly Choate and Miller (1992) describe the process of CBA as beginning with an extensive examination of the curriculum in question.

Another common guideline relates to the selection of writing samples for evaluation. Silliman and Wilkinson (1994) highlight the importance of using representative samples of a student's work. One recommendation is the use of portfolios which contain examples of a student's best writing as the basis for CBA (King-Sears, 1994). Howell, Fox, and Morehead (1993) suggest the use of already existing writing samples as the basis of CBA, as prior knowledge and interest in a writing topic are critical for good writing to occur.

Several aspects of writing may be evaluated using CBA approaches. Isaacson (1991) describes procedures for measuring writing fluency, syntactic maturity, vocabulary, content, and writing conventions. Poteet (1992b) describes procedures for evaluating handwriting, spelling, mechanics, usage, and ideation. King-Sears (1994) also describes several CBA procedures for evaluating letter formation, spelling, and sentence and paragraph writing. One frequently cited measure used in CBM is writing fluency (Howell, et. al., 1993; Isaacson, 1991; King-Sears, 1994; Marston, 1989; School District #57, 1996).

Measuring writing fluency. Currently, writing fluency is one of several forms of curriculum-based measurement (CBM) being used in School District #57 Prince George. Locally normed CBM is used to evaluate writing skills (School District #57, 1996) by having a student write for three minutes from a story starter. Scores derived from this instrument include the total words written (TWW) and the number of words spelled correctly (WSC).

Caution should be taken when evaluating written language in this way. While Howell, et. al. (1993) suggest that story starters are useful for generating writing samples when classroom generated samples are unavailable, they caution that when the purpose of writing is not generated or perceived by the writer, the product may be compromised. The result may be a writing sample that is not representative of a writer's usual work.

Another caution for the use of the CBM writing probes described here relates to how they are used. These types of measures were not designed to be used as a substitution for other types of assessment (Canter & Marston, 1998; School District #57, 1996). It is stated by School District #57 that "Curriculum Based Measurement provides only one of several pieces of required information. By itself, it is insufficient information" (p. 2). This is particularly true as this method of writing evaluation focuses on speed and spelling while ignoring other areas of writing such as those suggested by Isaacson (1991), King-Sears (1994), and Poteet (1992b).

Several studies report the reliability of these CBM writing measures (Marston & Deno, 1981; School District #57, 1996; Tindal, Marston & Deno, 1983). These studies found inter-scorer reliability ranging from .90 to .98. Marston and Deno found split-half reliability ranging from .96 to .99. In a measure of internal consistency comparing each minute of the writing to the other minutes using Chronbach's alpha, reliability ranged from .70 to .87. These were interpreted as satisfactory values for internal consistency. Measures of stability and equivalence conducted in the norming project by School District #57 at three month intervals revealed median coefficients of .62 and .67. Two studies of comparability of forms found reliability coefficients for of .73 and .95 for TWW (Marston & Deno, 1981; Tindal et al., 1983). In summary, Tindal et al. state that the findings of this research are that the procedures utilized in the CBM described here are generally reliable.

CBM measures have also been shown to demonstrate criterion-related validity. In a compilation of research findings on the validity of CBM (Marston, 1989), correlations ranging from .45 to .92 were found when comparing TWW and WSC scores on CBM to scores obtained on other measures of writing including standardized testing. In a study by Deno, Marston, and Mirkin (1982), CBM measures of WSC and TWW were compared to scores on the Test of Written Language (TOWL), the Word Usage subtest of the Stanford Achievement Test: Intermediate I, and the Developmental Sentence Scoring (DSS) System. Correlation coefficients ranged from .67 to .76 for WSC and .62 to . 84 for TWW. The highest correlations were for DSS and the Written Language Quotient on the TOWL. Using a multitrait-multimethod analysis, Tindal and Nolet (1990) found high discriminant validity between CBM and the Stanford Achievement Test measures of writing. They reported a correlation of .88 between scores on two CBM writing probes and the writing score on the Stanford Achievement Test. Fewster (2000) found that CBM writing scores obtained in Grades 6 and 7 were correlated to teacher-assigned letter grades in Social Studies and English in Grades 8, 9 and 10. Correlations demonstrated significant ($\underline{p} < .005$) small to medium effect sizes.

There is also some evidence favoring the discriminative validity of CBM writing measures. Tindal and Parker (1991) found that CBM measures of TWW and WSC were successful in detecting differences between children receiving specialized services in education in Grades 3 through 5, as compared to children not requiring this kind of support. This provides evidence that CBM scores can discriminate among students at different skill levels. Another recent study investigated the discrimination ability of CBM writing scores as well as their predictive validity. A discriminant analysis found significant differences between the CBM writing scores of children in each of the following groups: students in special education placements or receiving remedial support, students in regular education not receiving support, and honors students (Fewster, 2000).

Although these measures have been shown to demonstrate concurrent and predictive criterion-related validity and to discriminate among writers of varying abilities, the content validity of these measures is questionable. Content validity of a test is shown when the content is "drawn from the relevant environmental demands, that is, what the student is expected to do within the general education context" (King-Sears, 1994, p. 11). As writing from a story starter for three minutes is not a typical educational activity, it may be difficult to make generalizations about an individual's writing ability from performance on CBM writing samples to writing in general.

Measuring syntactic complexity. Another area for evaluation frequently suggested in the literature is that of syntactic complexity (Hughes et. al., 1997; Isaacson, 1991; Rousseau, 1990). All three authors suggest the T-unit as a useful measure for scoring syntactic complexity in writing samples. One T-unit, or terminable unit as named by Hunt (1965), consists of a main clause and any attached subordinate clauses. Hunt's study is frequently cited in the literature, and the T-unit continues as a common basis for measuring syntactic complexity in writing (Scott, 1988). Hunt studied the writing of children in Grades 4, 8, and 12. The writing samples she used consisted of 1000 words. All students in her study were of average intelligence. Her results indicated that T-unit length (mean length of T-unit or MLTU) and the ratio of clauses to T-units (subordination index or SI) were found to increase across the three grade levels. An analysis of variance using a factorial analysis showed that MLTU and SI were statistically significant ($p < .01$) for grade.

In an extensive longitudinal study concerning the spoken and written language skills of 211 children in Kindergarten through Grade 12, Loban (1976) also found a general increase in MLTU across grades. His findings for Grades 4, 8 and 12 were nearly identical to those of Hunt (1965). The growth in MLTU increased steadily in Grades 4 through 6 with some plateaus occurring in Grades 6 and 7.

Loban (1976) also studied the degree of subordination which he expressed as the number of subordinate clauses per sentence. These data were subsequently converted to the number of subordinate plus main clauses per sentence by Scott (1988) to allow for direct comparison to Hunt's work in this area. Like MLTU, the growth in SI increased between Grades 3 and 12, and again, the data for Grades 4, 8, and 12 were similar to the findings of Hunt (1965). Loban's study also made comparisons between low and high achieving students. His results showed higher MLTU scores for the higher group across all grades.

More recent studies also have investigated the usefulness of T-unit analysis for examining syntactic complexity. For example, Klecan-Aker and Hendrick (1985) found statistically significant differences between the T-unit lengths in the oral language of students in Grades 6 and 9 ($\underline{p} < .05$). There was not a statistically significant increase in the number of clauses per T-unit (SI) between the two grades. It should be noted that this study was conducted using oral language samples and while findings may not be generalizable to written language, it does provide more evidence that MLTU increases with growth in language development.

Summary. While these measures of writing fluency and syntactic complexity provide countable information on samples of writing, and have been well studied to establish their reliability and validity, there are limitations to their use. Again, like many commercial writing assessment tools, they focus on word and sentence level aspects of writing without taking into account larger discourse related aspects of writing such as cohesion and organization. The next section explores a class of tools that may be used to examine a variety of aspects of written language.

Categorical Assessment Tools

Silliman and Wilkinson (1994) suggest that categorical tools are useful for coding behaviors and skills quantitatively in language assessment. Two of the systems they suggest for accomplishing this include rating systems and checklist systems. Categorical tools will be examined from the perspective of their usefulness in evaluating cohesion in writing.

Rating systems. Rating systems may be used in a variety of ways to evaluate writing. Three kinds of rating systems are typically employed in writing evaluation. These are holistic ratings, analytic ratings, and primary trait ratings. Analytic and holistic ratings of writing are used in some of the standardized tests already mentioned. These procedures may also be applied to scoring writing samples produced in the classroom. Dagenais and Beadle (1984) described these procedures of evaluation as being helpful for use in classrooms and for planning instructional programs.

Holistic evaluations involve rating a writing sample on the basis of the overall presentation rather than its specific features (Miller, 1999). As seen previously, this type of rating does not allow for an in-depth analysis of specific areas of writing difficulty or success. Holistic evaluations require that the evaluator be trained in the use of the rating scale and are aimed at gaining an overall impression of a sample of writing. This impression can be gauged against a pre-established criteria (Dagenais & Beadle, 1984) or a group of writing samples may be rated according to their relative standing in relation to other writing samples (Miller, 1999).

Another type of rating consists of primary trait scoring. It involves examining particular aspects of a piece of writing and rating them individually (Sax, 1997). This type of scoring requires the development of criteria that are specific to the writing purpose. For example, a primary trait rating scale used to score an argumentative piece of writing could rate the persuasiveness of the argument. Students are rated against this criterion rather than against one another (Miller, 1999). According to Miller, this type of rating is more difficult to develop than other types due to the specific nature of the scoring criteria. However, due to the level of specificity, it provides more diagnostic information than holistic scoring.

A third form of rating consists of analytic scoring. This type of scoring involves rating different aspects of the same writing sample individually. For example, this type of scoring procedure may involve rating story development, grammar, and spelling each on a 5 point rating scale. An individual's score would be the total of all three ratings. According to Miller (1999), while having the advantage of analyzing different areas of writing strengths and weaknesses, this type of method is time consuming and may be impractical for large scale assessments. Furthermore, as found in the review of standardized writing assessments, this type of rating provides only a single measure for each skill area examined.

An example of an analytic rating scale is the writing reference set developed by the British Columbia Ministry of Education (1996b). This scale includes description or rubrics to assist teachers in scoring writing samples using a seven point scale on the features of Meaning, Style, Form, and Surface Features.

The rating systems described here generally are used to evaluate constructs globally and offer only a wide perspective analysis (Silliman & Wilkinson, 1994) of writing skills. Furthermore, these types of evaluations overall tend to be subjective in their scoring and can be time consuming to complete (Sax, 1997). Consequently, it is my opinion that such tools would not be the best means of providing a diagnostic measure of the various cohesive devices that children use in their writing.

Checklists. A checklist is another categorical system that may be used in the evaluation of writing. When evaluating writing, checklists can help focus the evaluator's attention on relevant details for scoring (Sax, 1997) rather than on making overall judgments about a construct, as occurs when using holistic or analytic rating procedures. While Silliman and

Wilkinson (1994) caution that checklists provide only a broad evaluation focus and may not be sensitive to small changes in communication behaviors, Sax indicates that checklists are useful in measuring complex behaviors that can be broken down into specific segments. Rousseau (1990) suggests that simple checklists are useful to pinpoint errors for error analysis and to allow for repeated direct measures of progress in writing development. Although evaluation with this kind of tool does not provide a qualitative look at a given behavior, it does allow for easy and inexpensive administration and comparison of a wide range of behaviors across a large number of students (Silliman & Wilkinson, 1994). It is the opinion of this author that the usefulness of a checklist for diagnostic purposes is probably related to the degree of specificity in the items. That is, broad items would allow for only broad assessment, whereas many specific items which reflected multiple aspects of a writing area could result in a fairly in-depth analysis of writing skills. Although a checklist will lose some information due to the absolute nature of rating only the presence or absence of aspects of cohesion, for example, this same trait increases the ease and objectivity of administration.

Given the arguments for the usefulness in using checklists for error analysis in diagnostic assessments, this method of evaluation seems the most viable for an analysis of the different markers of cohesion used in the writing of children. The other benefit to this form of assessment is its usefulness with a variety of curriculum-generated writing samples. The remainder of this discussion will focus on considerations for the development and evaluation of effective checklists.

Considerations for Checklist Development

One consideration in checklist construction involves a comprehensive analysis of the important aspects of a given behavior (Sax, 1997). According to Sax, construction of a checklist involves an in-depth knowledge of the skill to be evaluated. It is from this analysis and knowledge that the detailed content of the checklist is developed. Silliman and Wilkinson (1994) remind us that items on a checklist should be specific.

Regardless of the purpose of an assessment or the type of evaluation tool to be examined, Tindal and Parker (1991) suggest several further considerations for test development and evaluation. According to these authors, tests should have a method of standardized administration and demonstrate reliable scoring. They should discriminate between students with varying skill levels and show at least low-moderate correlations with other acceptable methods of assessment. They also should be sensitive to improvements in student abilities. King-Sears (1994) advocates for standardized procedures and content for measures to ensure integrity and to avoid compromising reliability and validity.

Reliability of Checklists

If measurements are to be reliable, scorer reliability is a must as scorer reliability places an upper limit on the reliability of the overall measure (Sax, 1997). Tindal and Parker (1991) state that "Clear and standardized administration and inter-scorer reliability are necessary for others to unambiguously interpret the results" (p. 211). One factor that affects the reliability of ratings on checklists is the ambiguity in the definitions of the trait to be measured. Other factors include the differences among raters that relate to training in the use of the instrument or the tendency of individual raters to score leniently or too severely on a consistent basis (Sax, 1997). It follows then that in developing a checklist, item clarity, rater training, and interrater agreement should be areas of focus.

Establishing Validity of a Checklist

There are several ways of viewing the validity of an evaluation tool. Content validity is established when the skills outlined in the instrument's items correspond to the skills one is claiming to assess. As mentioned earlier, the development of a checklist involves a comprehensive analysis of skills to be developed. One approach to establish content validity is by reviewing studies of how cohesion is used in writing and developing checklist items directly from the findings of research.

Another form of validity, called criterion-related validity, comes from concurrent writing assessments. That is, concurrent criterion related validity can be established by correlating the scores on an instrument to other related measures (Sax, 1997). As no other tests examining cohesion have been found, concurrent validity would have to be established by comparing scores on a cohesion checklist to scores on other assessments of writing skill. If cohesion is related to writing proficiency as indicated in the literature, then it should demonstrate positive correlations to other measures of writing proficiency.

Another form of validity, called construct validity, indicates the extent to which an assessment tool measures the theoretical construct it claims to evaluate. This form of validity includes many lines of evidence (Moss, 1995). Sax (1997) outlines several avenues that support an argument for construct validity. One line of evidence is the justification that the construct in question has educational relevance and importance. Another is that the construct can be measured. Convergent validity provides another line of evidence. This form shows multiple sources of evidence of the construct established through criterion-related and content-related validity arguments. Another argument for construct validity results from evidence of discriminant validity, that is evidence showing to what the construct is not related.

Thus, establishing the validity of an instrument involves multiple lines of evidence that can be established by looking at relationships between different measures and through examining the literature for explanations of the construct in question.

Research Purpose

As discussed in the foregoing chapters, there is a need for a tool to evaluate cohesion in the writing of school-aged children. Such a tool would be useful for tracking the development of cohesion in students' writing and for devising and monitoring written language intervention plans for students with writing difficulties. The purposes of the current research are two-fold:

1. To develop a checklist that can be used to evaluate cohesion in the writing of elementary school children.

2. To evaluate the reliability and validity of the checklist.

Scope of the Proposed Research

For the purposes of this study, cohesion will only be examined in one writing genre, narrative writing. The primary goal is the development of the items on the checklist with attention focused on creating a reliable and valid instrument. The development of scoring norms, considerations for developmental and cultural differences, and genre differences are topics for future research. Although the generalizability of this tool will likely be limited by the nature of the writing samples used in the development, it is felt that this research will provide a good starting place for the development of a tool that can be later extended to a wider variety of writers and types of writing.

Contributions of this Research

Given the limitations of currently used methods of assessment, and the impact of cohesion on writing quality, the research done here will contribute in three main ways. The first contribution is to the body of literature in the area of writing assessment. Hedberg and Fink (1996) indicate that the research on the writing of children with disabilities and story writing in general is sparse. This study will provide information that informs this area.

Another area to be informed is the body of research regarding cohesion. As found to be the case by this author, the information on cohesion use in writing in the research literature is sparse. Findings from this study will provide further information about cohesion in the narrative writing of elementary school-aged children.

The final area of contribution is to practitioners conducting writing assessments. Development of such a tool would help professionals detect and describe difficulties problem writers have in structuring written text (Lindeberg, 1984). Such a tool could then be used to plan and monitor interventions aimed at improving the use of cohesive devices in writing.

CHAPTER THREE: METHOD

Research Design

The current study was designed to develop and evaluate a checklist for measuring cohesion in writing. The development process involved several steps that were adapted from procedures outlined by Crocker and Algina (1986). The first steps that they suggested in constructing a test included identifying the primary purpose of the test, identifying the behaviors that represent the construct to be examined, preparing a set of test specifications, constructing an initial item pool, having items reviewed by knowledgeable panels followed by revisions as necessary, and preliminary item testing followed by revisions as necessary. These steps constituted the preliminary development of the instrument. The remainder of the steps they suggested included testing of items on a large sample that represents the population for whom the instrument is intended; determining the statistical properties of the item scores with elimination of items that do not perform as expected; conducting reliability and validity studies on the final form of the test; and developing guidelines for the administration, scoring, and interpretation of the instrument. These final steps formed the second part of this study, the large scale evaluation of the instrument. This chapter describes these steps in detail along with the data source used to conduct the item analyses, reliability and validity checks. This chapter describes both the statistical and qualitative procedures used.

Data Source

The data source used in this study consisted of 342 archival CBM writing samples from children in School District #57 in Grades 4, 5, 6, and 7. The samples were gathered from three elementary schools, and represented each school's entire Grade 4, 5, 6, and 7 population. A fourth school also provided samples but the set from this school was incomplete. As it was not certain that the samples represented the entire population of students in all four grades at this school, this set of writing samples was not included with the 342 used in the large scale study. However, I chose 20 writing samples from this fourth school to use in the preliminary item analysis.

As indicated in the previous chapter, the CBM writing samples used here are samples that are obtained by having children write for three minutes from a story starter. The samples are then scored for the total number of words written (TWW) and the number of words spelled correctly (WSC). The samples are gathered on a routine basis by many schools in School District #57 as one way of monitoring student progress. The three story starters are presented in Appendix B. Several examples of writing that reflect the range present in these samples are displayed in Appendix C.

I chose this kind of writing sample for use in this study for three reasons. One reason relates to the use of CBM in School District #57 where I work. Part of my aim in this study was to develop an instrument that would have practical use for myself and my colleagues. Because CBM use is prevalent in this district, developing an instrument that would be able to evaluate cohesion in writing using CBM writing samples would enable practitioners to capitalize on a resource already being used in the district. Furthermore, and more importantly, development of an instrument that worked with these writing samples would allow practitioners to extend the purpose and value of CBM beyond measures of fluency and spelling.

Another reason for choosing these samples relates again to the practical utility of the instrument being developed. If the cohesion checklist was able to measure differences in short writing samples, it would likely have utility for longer samples as well. The reverse, however, might not be true. That is, a checklist that could measure differences in the use of cohesive devices on longer samples may not be as sensitive to differences in shorter samples.

The final reason for choosing these samples is their similarity. In order to evaluate my checklist, I wanted to be sure that the variability in checklist scores reflected the performance of the checklist items. Therefore, as much as possible, I eliminated variability caused from sources such as differences in genre, audience, amount and type of instruction given for the writing task, or the amounts of supported editing and re-writing. CBM samples are generated with a standardized procedure and are available in large quantities across elementary grades. The procedures for administering CBM writing probes are presented in Appendix B.

The schools from which the writing samples were gathered were chosen on the basis of the availability of complete school sets of writing samples. To ensure variability in the data source, I used writing from each school's entire Grade 4, 5, 6, and 7 population, including writing by children with special needs, English as a Second Language/Dialect (ESL/D), and learning disabilities (LD). This constituted a convenience sample. As this research focused on developing checklist items rather than making generalizations about a population of students, random sampling was not necessary.

The writing samples collected had all identifying information such as the student's and school's name removed. I gave each writing sample an identification number and coded each one for grade, gender, and special learning designation. Special learning designations included ESL/D, LD, Special Learning Resource (SLR) which refers to children with IQ s of lower than 75, and Other which included children with behavior difficulties and hearing impairments. I also included a code to indicate which of three story starters was used. I also recorded scores for TWW and WSC.

Ethical Considerations

As the data used in this study was archival and contained no identifying marks, it was not necessary to obtain the consent of individuals. As this research focused on the evaluation of an assessment tool rather than students, there was no perceived harm to individuals. Norm Monroe, Director of School Services, provided written consent to conduct this study using writing samples from Prince George School District #57 (see Appendix A).

Procedures

Preliminary Development of the Instrument

Initial Compilation

I carried out the first steps of identifying the primary purpose of the test and behaviors that represent the construct to be examined, in this case cohesion, through review of the literature. These purposes and behaviors are described in Chapter Two. I developed each item on the instrument as well as the table of specifications from the compilation of research findings on cohesion as outlined in the literature review, and based largely on the definitions of cohesion developed by Halliday and Hasan (1976). The first steps I used in evaluating this instrument consisted of panel reviews, a preliminary item analysis and a pilot interrater study.

Panel Reviews

The first step in revising the checklist involved two panel reviews. The first panel consisted of myself and three teachers with experience in testing who had taken graduate courses in measurement and evaluation in education. This evaluation focused mainly on the structural aspects of the checklist. This included determining if the items were free from technical flaws such as errors in spelling and grammar; determining the accuracy, appropriateness, and relevance to test specifications; judging the level of readability; and examining item bias and ambiguity of items. A second panel consisted of myself and five other school speech-language pathologists. With this panel, discussion focused on how well the items on the instrument reflected the concept of cohesion. This group also supplied feedback on the clarity of items and examples, as well as the layout of the instruction manual and ease of scoring. Following feedback from these two panels, I made a number of minor revisions to the wording of items and the format of the instrument. These changes will be discussed in more detail in the Chapter Four.

Preliminary Item Analysis

In the next phase of development, I scored 20 writing samples using the cohesion checklist and performed a preliminary classical item analysis on the results. The 20 samples I selected for this portion of the study were not used in the large scale testing. This group of 20 samples include five from each of Grades 4 through 7. I selected each sample by "eyeballing" the overall length and legibility. I chose the first five I found that represented a mid-range length for the grade.

I then used the results from this scoring for an item analysis using ITEMAN (1994). Item analysis is a procedure for examining the statistical performance of items on a assessment instrument. The statistical results help to determine which items are too easy or too difficult, and how well items discriminate between high and low scorers. ITEMAN is a classical item analysis software program which calculates standard item statistics and summary statistics. I used the results from this analysis to "red flag" items that could show up as problems during the next phase of the preliminary development. These flagged items may be ones that show up as ambiguous in the interrater study. At this point I made some changes to the checklist items and a second version of the checklist was created. These changes are described in Chapter Four.

Pilot Study of Interrater Agreement

In the pilot study of interrater agreement, I and 12 volunteers scored ten writing samples for comparison of agreement among raters. Each rater scored the same ten writing samples. I selected these samples, by using a random numbers chart, from the 342 samples collected for the study . The volunteers consisted of three school speech and language pathologists, seven learning assistance or remedial teachers, one classroom teacher, and one school psychologist. This phase of the study was conducted over a day-long session.

Training session. The first half of the day consisted of a training session. Training consisted of an overview of the notion of cohesion, followed by a thorough review of the checklist and scoring practice. During the checklist review, I led the group of volunteers through an item-by item examination of the checklist, using examples to illustrate appropriate scoring of each item. The next step in training focused on practice with scoring writing samples. All participants scored the first two practice samples in small groups to allow raters the chance to discuss and clarify their scoring choices. They then scored the next two examples individually. In all cases of scoring practice, the group reconvened to compare scores and discuss differences. The raters completed two additional practice samples, as they expressed uncertainty in their scoring and there was still some disagreement on some of the items. By the end of the scoring practice, at least 11 out of 13 raters agreed on the scoring of each item on a given example. After this final practice, scoring of writing samples for the pilot interrater portion of the project commenced.

Scoring session. For the scoring portion of the session, I provided each rater with a bundle of eight writing samples. When scoring was completed on these, I supplied another bundle of eight. Each bundle included five copies of probes that would be scored by the whole group. The other 3 samples were different for all. I included the extra writing samples and used a random arrangement of interrater writing samples within the bundles to reduce the opportunity for raters to compare scores as it was unlikely that any two raters would be scoring the same writing sample at the same time. Also, this made the raters blind to which samples would be used for interrater comparison. Each rater scored 10 interrater probes in all.

Data from one rater was eliminated from the study. This rater was only able to complete five interrater probes and noted when handing them in that she had miscopied the probe identification numbers. Therefore it could not be determined which sample went with which completed checklist. One other rater returned only eight of the ten samples and another rater returned only nine.

Using the completed checklists from the 12 raters, I calculated the proportion of agreement for each writing sample on an item-by-item basis. I then calculated the mean proportion of agreement across all 10 interrater writing samples for each item. I also compared total scores on the checklists.

Revisions to Checklist Content and Process

Modification of items. The preliminary item analysis and pilot interrater studies were instrumental in eliminating major problems inherent in the checklist, both in content and scoring procedure, prior to carrying out the large scale study. It is noteworthy that discussions with and among the group during the training portion of the interrater pilot study led to many valuable suggestions about improvements that could be made to reduce ambiguity in the checklist items and scoring instructions. These suggestions included specific examples that would clarify when and when not to give credit on individual items. Each volunteer wrote notes and comments on their instruction manuals to assist them in scoring. I then collected these notes at the end of the session as additional input to consider when evaluating and editing items on the checklist.

I used the results of the preliminary item analysis in combination with the data on the proportion of agreement between raters on each item to delete and modify checklist items prior to the large scale study. While quantitative results are emphasized in this report, the process of deciding how and where to change items also involved qualitative processes.

Specifically, I used a reiterative approach in examining the literature on cohesion studies, examining items that performed poorly in the item analysis, examining notes taken during the rater training session and examining items which performed poorly in terms of proportion of interrater agreement. This process resulted in another revision of the checklist.

Procedural modifications. Not only did the pilot portion of this study lead to changes in checklist items before the final analysis, but it also raised questions about whether or not all writing samples could be adequately scored using the checklist. When looking at interrater agreement, it was clear that agreement overall was much lower on some writing samples than others. I examined these samples for qualities that might suggest criteria for inclusion in the study. This examination resulted in the following criteria for a sample's inclusion.

First, the sample had to be readable. That is, the handwriting and spelling patterns had to be such that words could be deciphered. This did not mean that spelling errors could not be present, but the words had to be decipherable. I chose a limit of 2 unreadable words as the cut-off criterion for inclusion. Second, the sample had to contain at least two sentences. The rationale for this criterion was based on the fact that the checklist was meant to analyze inter-sentence connections. If sentence boundaries were not present, cohesion could not be scored. I removed writing samples not meeting these criteria from subsequent phases of the study.

Evaluation Using a Large Scale Sample

The next part of the study involved scoring 312 of the 342 collected writing samples using the third revision of the checklist which contained 25 items. Thirty of the writing samples had been eliminated from this part of the study as they did not meet the criteria for inclusion. Seventy Grade 4 students, sixty-seven Grade 5 students, eighty-four Grade 6 students and eighty-nine Grade 7 students generated the samples used. Two writing samples missing the code for grade were retained in the study but were eliminated from analyses conducted by grade. Of these remaining writing samples, 30 were written by students designated ESL/D, 7 by students designated SLR, 21 by students designated LD and 9 by students designated as 'other'. I scored all samples in this portion of the study for cohesion, the number of T-units, mean T-unit length (MLTU), and the degree of subordination (SI). I then used this scoring of the writing samples to perform several analyses on the checklist items. After additional revisions of the checklist items, I conducted another interrater study. I also used the scores obtained in this portion of the study to establish concurrent criterion-related validity.

Item Analysis

An item analysis was run using ITEMAN (1994). I chose a classical item analysis to conduct this study as it provides statistics that reflect the performance of items including information about their discrimination. A Rasch analysis would also provide information appropriate for analyzing item performance but does not provide information about the discrimination ability of items. This method assumes that all items discriminate equally (Sax, 1997). Rasch scales also assume that the latent trait being measured is unidimensional (Sax, 1997). As this instrument was a new creation, I did not know if these assumptions could be met.

As well as providing statistics that reflected item performance, this classical analysis provided a measure of internal consistency for the overall checklist. I deleted or combined items that performed poorly on this analysis and ran the analysis a second time to determine the discrimination of items on this last version of the checklist which now consisted of 13 items. I also ran a third item analysis on these 13 items to determine how items performed as parts of subtests rather than as parts of the total checklist.

Interrater Study

Another step in the large scale study was to establish the reliability of the instrument through an examination of interrater agreement and reliability. Tinsley and Weiss (1975) indicate that it is important to gather indices of both interrater agreement and reliability. They indicate that agreement is established by examining the extent to which different raters give the same scores on an instrument. This can be reported in terms of proportion of agreement. Interrater reliability, on the other hand, indicates the degrees to which scores from different raters are proportional to one another. This is usually reported in terms of indices of analysis of variance and correlational values (Tinsley and Weiss, 1975).

I completed this portion of the study with three other raters. My scores were excerpted from the large scale study. In this interrater study all four raters were school speech-language pathologists. This professional group shares a common background with respect to clinical measurement of children's language, and represents the practitioners that would most likely use such an instrument. For this interrater study, I selected probes consisting of only one story starter. The samples were also marked for sentence boundaries to eliminate the need for judgment in this regard as not all raters were familiar with the procedures to do this. By taking these steps, it was easier to interpret the variability among raters' scores as reflecting problems with the checklist rather than some other variable. To ensure adequate variability in the scores of writing samples used in this interrater study, I chose samples that represented the full range of scores of 1 through 12 on the checklist.

Each rater received copies of the checklist and manual, as well as two practice items in advance of the training session. The training session was scheduled for a half day. During this time, I focused on training the group to recognize examples and non-examples of checklist items in relation to the rules for scoring outlined in the manual. Several practice samples were then scored to establish understanding of the scoring rules with the raters. At the conclusion of the training session, I presented the raters with a package of 12 probes to score independently.

Once the completed checklists were collected, I examined the results for the proportion of agreement among raters on an item-by-item basis. I then calculated a one-way analysis of variance (ANOVA) and Levene's F statistic to determine if any significant differences existed between raters. I also calculated correlations for agreement between rater pairs.

Validity Measures

In this portion of the study I compiled several sources of evidence for the validity of the checklist. The content validity of this instrument was supported by the review of the literature presented in Chapter Two. Chapter Two also contains arguments for the educational relevance and importance of measuring the construct of cohesion. This argument supports the construct validity of this instrument. To provide further evidence of construct validity, I conducted a factor analysis on the items. I also conducted a discriminant analysis to determine the predictive value of checklist scores for grade membership.

The final area of validity that I investigated in this study was concurrent criterion-related validity. As no other measures of cohesion were available, I turned to other measures of writing that related to writing proficiency. I used measures of writing fluency and measures of syntactic complexity calculated on the same writing samples from which I obtained my cohesion scores. Fluency measures consisted of total words written (TWW) and words spelled correctly (WSC). Syntactic measures consisted of the mean length of T-unit (MLTU) and the subordination index (SI). These terms were defined in Chapter Two.

Writing fluency measures. I chose these scores for concurrent criteria as the reported findings (see Chapter Two) indicated that CBM measures are reliable and relate to other overall measures of writing. Performance on these instruments has been shown to relate to other measures of writing development, and to discriminate between students with learning problems and those without. It was therefore expected that CBM scores would provide a reasonable criterion of writing proficiency.

As cohesion has been shown to relate to overall writing proficiency, scores on a checklist of cohesion should also relate to a variety of measures that also relate to writing proficiency. If scores of TWW and WSC are related to overall writing ability as implied in the literature reviewed in Chapter Two, then they should relate to measures of cohesion.

Syntactic complexity measures. Syntactic complexity is a relevant measure of writing proficiency because research has long shown that syntactic complexity and elaboration are clear measures of writing development (Hunt, 1965; Loban, 1976). Furthermore, studies of "problem writers" show that their products contain sentences which are less complex (Ratner & Harris, 1994) especially in regards to the degree of subordination (Scott, 1991b). Many other researchers have also found differences between problem writers and "good writers" in both length of sentences and syntactic complexity (Anderson, 1982; Poplin, Gray, Larsen, Banikowski & Mehring, 1980; Scott, 1991b; Singer, 1995). Therefore, it can be argued that poor writers are likely to score poorly on measures of syntactic complexity while good writers should score better. Consequently, it is expected that measures of MLTU and SI will show positive correlations to other measures of writing ability. Assuming that scores on a measure of cohesion are related to writing ability, correlations between MLTU, SI and scores of cohesion are expected.

Prior to using measures of TWW, WSC, MLTU and SI for concurrent validity purposes, I examined their statistical performance with this data source to further determine their effectiveness as criterion measures. This included calculations of means and standard deviations of each of these scores. I then ran a discriminant analysis to determine the ability of these measures to predict grade. Finally, I calculated correlations between checklist scores and scores of TWW, WSC, MLTU and SI to determine the concurrent criterion-related validity of the checklist scores.

Summary

The development and evaluation of a checklist to assess cohesion in writing involved many steps utilizing 312 CBM writing samples as the data source. The preliminary development of the checklist involved the compilation of checklist items, panel reviews, a preliminary item analysis, and a pilot interrater agreement study. The evaluation of the checklist involved classical item analysis; reliability checks for internal consistency and interrater agreement; and validity checks including a factor analysis, a discriminant analysis, and a concurrent criterion-related validity study.

CHAPTER FOUR: RESULTS

This chapter presents the results of the checklist development and evaluation procedures. The first section describes the outcome of the preliminary development of the instrument including results of the panel reviews, the preliminary item analysis, and the pilot study for interrater agreement. The next section reports the results of the large scale sample evaluation of the checklist. This section includes the outcomes of three item analyses as well as results of validity and reliability checks. The results of a second interrater study are also included here.

Preliminary Development of the Instrument

Initial Compilation of Items

Appendix D contains the preliminary draft of the checklist that I compiled based on the research literature. I will call this version Checklist 1.0. This was the form of the checklist presented to two panels for review.

Panel Reviews

The majority of comments from the first panel review focused on ambiguity in the wording of the items, overlap of item content and the weighting of items relative to the table of specifications. The second panel focused its attention on the ease of scoring and the clarity of examples and instructions. I considered the feedback from both panel reviews in conjunction with the aim of my research and that of the instrument.

As a result of these reviews, I made adjustments to some of the items. These included minor changes in the wording of some items, and a switch in the order of presentation of two items to clarify the scoring procedure. The panels also flagged some additional items as potentially problematic, but I left these unchanged pending the outcome of the item analysis.

I also changed the format of the manual based on suggestions made by the panels.

These changes included the addition of a brief explanation of cohesion in the introduction to the manual and inclusion of a section defining key terms. I also developed a scoring companion that provided point-form scoring explanations and examples for quick reference.

The panel reviews concluded with a second version of the checklist called Checklist 1.1. See Appendix E for a copy of this version of the checklist. Checklist 1.1 was used in the preliminary item analysis.

Preliminary Item Analysis

Table 2 shows the results of the classical item analysis conducted on Checklist 1.1 using 20 scored writing samples. Item analysis is defined as the computation and examination of the statistical properties of responses to individual test items to permit the selection of items on an instrument (Crocker & Algina, 1986). The ITEMAN (1994) analysis program used in this study generates three main item statistics.

The first one, proportion correct, indicates how many writing samples received credit on each checklist item. This indicates the difficulty of the item. Values above .85 indicate easy items while values at or below .49 indicate difficult ones (Sax, 1997). Items that are either too easy or too difficult are not desirable to retain on an assessment instrument because they provide little information about the individuals being evaluated. That is, if everyone is scoring the same on an item, the item does not discriminate among writers.

Table 2

Item Statistics from the Preliminary Item Analysis

Item	Prop.	D	r_{pb}	Item	Prop.	D	r_{pb}	Item	Prop.	D	r_{pb}
No.	Corr.			No.	Corr.			No.	Corr.		
1	.65	.46	.48*	10	.00	.00	-	19	.25	.09	.35
2	.35	.71	.54*	11	.30	.43	.26	20	.05	.00	-.05
3	.05	.14	.46*	12	.15	.43	.38	21	.00	.00	-
4	.65	.66	.44*	13	.15	-.06	.17	22	.25	.37	.35
5	.55	.46	.28	14	.00	.00	-	23	.20	-.26	-.16
6	.25	.37	.35	15	.25	.43	.43	24	.50	.03	.15
7	.90	.40	.51*	16	.80	.26	.11	25	.95	.20	.22
8	.50	.57	.41	17	.30	.23	.34	26	.00	.00	-
9	.00	.00	-	18	.55	.31	.35	27	.00	.00	-

Note. Item No. = item number; Prop. Corr. = proportion correct; D = discrimination index; r_{pb} = Point Biserial; - = data not calculated.

* $p < .05$.

Another item statistic, the discrimination index, indicates "the extent to which items differentiate between those persons with highest and lowest scores on the Total test" (Sax, 1997, p. 240). Positive discrimination values show that more examinees who scored high on the instrument have received credit for that item than examinees who scored low. Values between 0.31 and 1.0 indicate good discrimination. Values between 0.10 and 0.30 indicate fair discrimination, while values of 0.09 and below indicate poor discrimination (Sax, 1997).

Another statistic calculated by this program is the point biserial. This statistic shows how well the score on an item relates to overall performance on the instrument (Crocker &

Algina, 1986). The same approach to examining any correlation is applied to the examination of point biserial correlations. Both the size of the relationship and the statistical significance require evaluation.

From the table it can be seen that 11 of the 27 items showed poor discrimination on this preliminary item analysis. Of these, 6 items discriminated poorly because none of the writing samples received credit for these items. Of the remaining 5 poorly discriminating items, 2 (items 13 and 20) also had low proportions correct. The remaining 3 items (items 19, 23 and 24) were considered problematic due to nonsignificant point biserial correlations ($\underline{r}_{pb}$ (18) < .38, $\underline{p}$ < .10) and low or negative discrimination values that could not be explained by low proportions correct. The majority of the items demonstrated point biserial values that were not statistically significant ($\underline{p}$ > .05). I attributed this as probably due to the small sample size (n = 20) used in this portion of the study. As the results were only used as a preliminary indicator of potentially problematic items, I did not give great weight to the level of statistical significance at this stage.

Although these results were interpreted cautiously due to the small sample size, they prompted me to make some changes to the checklist. These changes reflected the results of the preliminary item analysis as well as notes from the panel reviews. In particular, I changed items 24 through 27 significantly.

The changes to these items can be summarized as follows. Checklists 1.0 and 1.1 contained three items that evaluated the sophistication of organization achieved through the use of paragraph or paragraph-like structures. After the preliminary item analysis, these items were collapsed to a single item with the content of the revised item reflecting the kinds of paragraph structures more likely to be found in a narrative writing sample. In addition, I added two more items to the checklist that evaluated lexical cohesion in a more specific manner.

These changes resulted in a new version of the instrument, Checklist 1.2. The reader is referred to Appendix F to review this version of the checklist.

Pilot Study for Interrater Agreement

In this interrater study, 12 raters scored 10 interrater writing samples using Checklist 1.2. The interrater samples were randomly chosen from the data source of 342 writing samples collected.

Table 3 displays the mean proportion of agreement among raters for each item across writing samples. I arrived at these figures by calculating the proportion of ones or zeros scored for a given item across all 12 samples. The table shows a large degree of variability in the scoring of items as indicated by varied means and large standard deviation values for these proportions. Agreement ranged from 69.02 percent to 100 percent. I interpreted this variability in scoring as reflecting ambiguities in some of the checklist's items. I considered items with less than 85% agreement among the raters to be problematic. This proportion of agreement was chosen as a cut off as this level indicated that a mean of 11 out of 13 raters agreed on the scoring of that item across writing samples. As this would indicate a majority of raters agreeing, this was felt to be an indicator that the checklist item was not ambiguous.

Table 3

Mean Percentage Agreement Among Raters Across Items

	% Agreement		% Agreement		% Agreement
Item No.	M (SD)	Item No.	M (SD)	Item No.	M (SD)
1	83.18 (12.43)	10	96.59 (5.90)	19	95.76 (4.48)
2	72.36 (13.20)	11	98.33 (3.51)	20	95.83 (10.58)
3	87.20 (17.85)	12	100.00 (0.00)	21	97.42 (4.15)
4	75.98 (18.58)	13	96.59 (8.11)	22	93.11 (11.41)
5	81.52 (17.91)	14	95.76 (10.61)	23	92.42 (12.08)
6	69.02 (11.31)	15	83.18 (17.83)	24	93.26 (6.60)
7	91.36 (14.48)	16	94.77 (7.36)	25	83.03 (14.71)
8	93.11 (12.69)	17	91.09 (10.61)	26	96.59 (5.90)
9	99.17 (2.63)	18	84.45 (13.28)	27	95.00 (10.54)

Note. Item no. = Checklist item number; M = mean; SD = standard deviation.

The results of this interrater study indicated that items 2, 4 and 6 had particularity low levels of agreement. Items 1, 5, 15, 18 and 25 were also considered problematic by the results indicated here.

Revisions to Checklist Content and Process

I made several changes to the checklist as a result of the preliminary item analysis and pilot interrater studies. I considered the results from these sections in conjunction with feedback from the panels and raters, as well as from the findings of research reviewed in the literature. Changes I made included combining, deleting, and expanding some items. This was done to reduce ambiguities, to eliminate redundancies, and to reduce the number of items that were either too difficult or too easy. Other changes included adding definitions and examples to the instruction manual and scoring companion. I also made some changes to the format of the checklist such as dividing it into subtest sections for ease of scoring and improved appearance. In addition, I added a section to the manual describing the method of preparing a writing sample for scoring. This included criteria for writing samples with which the checklist may be used and instructions to use T-units rather than sentences as the basis for analysis. The result of these combined findings was another version of the checklist containing 25 items, called Checklist 2.0. See Appendix G for a copy of this version of the checklist.

Evaluation Using a Large Scale Sample

This section includes the results generated after I scored all 312 writing samples that met the criteria for inclusion in the study, using Checklist 2.0. Of the 30 samples that did not meet criteria for inclusion, thirteen were from Grade 4, nine from Grade 5, five from Grade 6 and three from Grade 7. Twenty three of the excluded samples were generated by children with special designations. Students designated ESL/D wrote 6 of the excluded samples, students designated as SLR wrote 4 of the excluded samples, students designated LD wrote 11 of the excluded samples and students designated as "other" wrote 2. I included several samples of writing used in this portion of the study in Appendix C. These include good and poor examples of writing as well as typical samples for each grade.

The information reported in this section includes three classical item analyses using the ITEMAN (1994) software. In addition, I report the data that provide evidence for checklist validity and reliability. The results of the final interrater study are also included in this section.

Item Analysis

The mean total score on the 25 item checklist was 7.69 with a standard deviation of 2.20. The minimum score was 2 and the maximum was 16 with a median of 8.

Results of the item analysis are displayed in Table 4. All items had positive discrimination indices. Items with poor discrimination indices and low point biserial correlations were considered for revision. I eliminated two items (21 and 25) from the checklist due to poor discrimination. I removed another two items (22 and 23) as they were too easy. Item 24 was also deleted as it was the only other item remaining on that subtest. I combined remaining items with low proportions correct with other items of related content.

Table 4

Item Statistics from the Analysis of Checklist 2.0

Item	Prop	D	r_{pb}	Item	Prop	D	r_{pb}	Item	Prop	D	r_{pb}
No.	.			No.	.			No.	.		
	Corr.				Corr.				Corr.		
1	.66	.36	.32***	10	.12	.18	.31***	19	.08	.11	.13*
2	.51	.39	.36***	11	.03	.03	.06	20	.62	.31	.32***
3	.20	.21	.27***	12	.09	.09	.18**	21	.07	.09	.15**
4	.55	.54	.46***	13	.27	.31	.33***	22	.97	.07	.15**
5	.25	.16	.16**	14	.33	.44	.14***	23	.91	.16	.24***
6	.50	.33	.30***	15	.05	.06	.09	24	.65	.36	.31***
7	.01	.01	.10	16	.01	.03	.10	25	.02	.04	.13
8	.02	.01	.00	17	.29	.27	.30***				
9	.39	.15	.14*	18	.09	.13	.26***				

Note. * $p < .05$. ** $p < .01$. *** $p < .001$.

Taking into consideration the deleted and combined checklist items, the scoring results were adjusted to produce another composition of items. This composition formed Checklist 2.1. I then ran another item analysis to evaluate Checklist 2.1. This form of the checklist consisted of 13 items that were divided across three subtests called Reference, Conjunction and Lexical Cohesion. This version of the checklist is displayed in Appendix H. On this classical item analysis the mean total score on the checklist was 4.92 with a standard deviation of 1.92. Mean scores (SD) for the individual subtests were 2.16 (1.24), 1.99 (1.18) and 0.77 (0.62) for Reference, Conjunction and Lexical Cohesion respectively.

The minimum score for Checklist 2.1 was 0 while the maximum was 12 with a median of 5. The individual item statistics are displayed in Table 5. The new combination of items improved the proportion correct on almost all items on the checklist.

Table 5

Item Statistics from the Item Analysis of Checklist 2.1

Item	Prop.	D	r_{pb}	Item	Prop.	D	r_{pb}
No.	Corr.			No.	Corr.		
1	.66	.36	.39***	8	.13	.14	.22***
2	.51	.46	.45***	9	.57	.29	.32***
3	.20	.20	.26***	10	.37	.34	.36***
4	.55	.50	.47***	11	.29	.21	.29***
5	.25	.25	.28***	12	.14	.21	.25***
6	.52	.24	.28***	13	.62	.37	.35***
7	.12	.18	.33***				

Note. *** = $p < .001$

The discrimination index on all but two items now fell between .20 and .50. Point biserial correlations for all items ranged from .22 to .45 (significant to $p < .001$).

A final item analysis was conducted on the 13 item form (Checklist Version 2.1) to compare individual item performance to other items in the same subtest rather than the Total

Test. These results are displayed in Table 6. On this run, the discrimination indices for all but two items now ranged from .21 to .93. Nine of the items now had a discrimination index over .40. Point biserial correlations showed similar improvement with values now ranging from .25 to .82 (significant to $p < .001$). The performance of almost all checklist items was improved by evaluating them as components of subtests rather than of the total checklist.

Table 6

Item Statistics from the Item Analysis Run Using Subtests

Subtest	Item	Sub	Prop.	D	r_{pb}	Subtest	Item	Sub	Prop.	D	r_{pb}
	No.	Item	Corr.				No.	Item	Corr.		
REF	1	1-1	.66	.72	.64	CON	8	2-3	.13	.18	.27
	2	1-2	.51	.88	.72		9	2-4	.57	.56	.50
	3	1-3	.20	.21	.25		10	2-5	.37	.60	.54
	4	1-4	.55	.44	.44		11	2-6	.29	.51	.49
	5	1-5	.25	.54	.59	LEX	12	3-1	.14	.22	.62
CON	6	2-1	.52	.63	.54		13	3-2	.62	.93	.82
	7	2-2	.12	.17	.31						

Note. REF = Reference Subtest; CON = Conjunction Subtest; LEX = Lexical Cohesion Subtest; Item No. = checklist item number; Prop. Corr. = proportion correct; D = discrimination index.

The only exceptions were items 3 and 7 which showed little or no improvement with this analysis.

Scale inter-correlations were also calculated. Correlations between the Reference subscores and the Conjunction and Lexical Cohesion subscores were r = .03 and .10

respectively. The correlation between the Conjunction and Lexical Cohesion subscores was $\underline{r} = .06$. These values indicated no relationship among the three subscales of the checklist.

A Pearson's $\underline{r}$ was calculated between the subtest scores and the Total Test scores. Correlations of subtest scores to the Total Test score were $\underline{r} = .70$, .67, .42 for Reference, Conjunction, and Lexical Cohesion subscores respectively. All correlations were significant ($\underline{p} < .001$).

Checklist Reliability

The internal consistency of the checklist overall was $\alpha = 0.32$ with a standard error of measurement (SEM) of 1.58. The internal consistency of subtests were $\alpha = .39$, for Reference, $\alpha = .22$ for Conjunction, and $\alpha = .10$ for Lexical Cohesion.

Interrater study. This section reports the results of the interrater study conducted with the 13 item Checklist 2.1 using myself and three other raters. The proportion of agreement for total scores was not reported as there were only four raters. Proportions would therefore only reflect agreement levels of 0, 25, 50, 75 and 100 percent. I felt that these increments were too large to be discerning. I did, however calculate the proportion of agreement on an item-by-item basis across writing samples to determine on which items raters most frequently disagreed. To establish interrater reliability, I calculated correlations between pairs of raters and ran a one-way analysis of variance (ANOVA).

Means and standard deviations were calculated for the 12 interrater checklist scores generated by each rater. The results showed mean scores of 5.75 ($\underline{SD}$ =2.13), 5.83 ($\underline{SD}$ = 2.14), 4.75 ($\underline{SD}$ = 1.85), and 6.00 ($\underline{SD}$ = 2.67) for raters one, two, three and four respectively. I was rater number four. While mean scores and spread of scores are similar for raters one, two and four, rater three's scores were slightly lower with less variability indicating more stringent marking by rater three.

Table 7 displays Pearson product moment correlations calculated between pairs of raters. The correlation coefficients between raters ranged from .70 to .91 and were significant to p < .05. The three lowest correlations all involved rater three. Coefficients between raters one, two and four ranged from .86 to .91.

Table 7

Correlations Between Rater Pairs Across Items

Raters	1	2	3	4
1	-	.86 (.000)	.75 (.005)	.88 (.000)
2	-	-	.70 (.011)	.91 (.000)
3	-	-	-	.71 (.009)

Note. p value in parentheses

Levene's test of homogeneity for variance among raters' scores generated an F statistic of .507 (3, 44) at a significance level of p = .68 revealing that the assumption of homogeneity was met. As can be seen in Table 8, a one-way ANOVA revealed no significant difference between raters.

Table 8

Analysis of Variance for Between Rater Differences

	Sum of Squares	df	Mean Square	F	p
Among Raters	11.500	3	3.833	.479	.699
Within Raters	352.167	44	8.004		
Total	363.667	47			

The examination of the proportion of agreement on individual items across samples showed which items may still be ambiguous and therefore causing lower agreement. As you can see in Table 9, agreement among raters across checklist items ranged from 75.00 to 97.92 percent. I considered percentages below 80 on a single item to be problematic. I selected this criterion as it indicated that a mean of more than three out of four judges agreed on the score of that item across all 12 checklists. Two items (items 5 and 13) showed less than 80 % agreement.

Table 9

Proportion of Agreement Across Writing Samples for Each Item

Item	% Agreement		Item	% Agreement		Item	% Agreement	
No.	M	(SD)	No.	M	(SD)	No.	M	(SD)
1	87.50	(16.85)	6	87.50	(19.94)	11	95.83	(9.73)
2	83.33	(12.31)	7	97.92	(7.22)	12	93.75	(15.54)
3	85.42	(22.51)	8	93.75	(15.54)	13	75.00	(21.32)
4	85.42	(16.71)	9	87.50	(19.94)			
5	79.17	(20.87)	10	95.83	(9.73)			

Note. M = mean; SD = Standard deviation.

Validity Measures

The validity measures reported here include the results of a factor analysis and a discriminant analysis. They also include results of the concurrent criterion-related study.

Factor analysis. I attempted a principal component analysis using the 13 items from Checklist Version 2.1. This was done to determine whether the checklist items loaded into components reflecting the instrument's subtests. A principal component extraction followed by Varimax rotation generated six factors with eigenvalues greater than 1. This model accounted for only 59.83% of the total variance. A three component solution with Oblimin rotation and Kaiser normalization was also attempted, but this solution only accounted for

35.30% of the total variance. The three factor model accounted for all variables with loadings ranging from -.301 to .808. The loadings for each component are displayed in Table 10.

Table 10

Variable Loadings for a Three Component Model

	Component				Component		
Item No.	1	2	3	Item No.	1	2	3
1	.663			8		.592	
2	.808			9			.448
3			.310	10			.721
4		.576		11			.559
5	.723			12		.424	
6	-.301			13		.456	
7		.467					

To determine why the factor analysis accounted for a limited amount of variance, I examined item-by-item correlations. These are displayed in Table 11. Tabachnick and Fidell (1996) indicate that in order for a factor analysis to work, the correlation matrix should include "sizable" correlations. They define sizable values as greater than .30. As can be seen by examining Table 8, very few inter-item correlations were present. Only two values exceeded .30. All reported r values were significant (p< .05).

Table 11

Matrix Displaying Item-by-Item Correlations

Item No.	1	2	3	4	5	6	7	8	9	10	11	12
2	.45											
3	-	-										
4	-	-	-									
5	.19	.44	-.12	-								
6	-	-	-	-	-.19							
7	-	.12	-	.13	.11	-						
8	-	-	-	.11	-	-	.10					
9	-	-	-	-	-	.18	-	-				
10	-	-	.13	.13	-	-	-	-	-			
11	-	-	-	-	-.12	-	-	-	-	.21		
12	-	-	-	.10	-	-	-	-	-	-	-	
13	-	-	-	.17	-	-	-	-	-	-	-	-

Note. - = *r* values both nonsignificant and less than .10.

Discriminant analysis. I also ran a discriminant analysis to determine if checklist scores could predict grade membership or special learning designation. I first calculated descriptive statistics for each grade's scores on the checklist. As can be seen in Table 12, although there was a general growth trend across grades on the total checklist score, there was little difference in the scores for Grades 4, 5 and 6. Table 13 shows more scattered patterns of growth with the subtest scores. Only Lexical Cohesion shows a steady improvement of scores across grades.

Table 12

Statistics Describing Checklist Total Scores by Grade

Grade	n	Range	Median	M	SD	SEM
4	71	0 - 9	4	4.46	1.86	0.22
5	67	0 - 10	5	4.76	1.35	0.20
6	84	1 - 9	5	4.67	1.10	0.21
7	89	1 - 12	6	5.64	0.58	0.23

Note. M = mean; SD = standard deviation; SEM = standard error of measurement

Table 13

Statistics Describing Checklist Subscores by Grade

Grade	Subtest	M	SD	SEM	Subtest	M	SD	SEM
4	REF	2.16	1.35	0.16	LEX	0.58	0.58	0.07
5	REF	1.97	1.18	0.14	LEX	0.69	0.61	0.07
6	REF	2.18	1.28	0.14	LEX	0.70	0.62	0.07
7	REF	2.30	1.18	0.13	LEX	1.10	0.69	0.07
4	CON	1.73	1.10	0.13				
5	CON	2.00	1.10	0.14				
6	CON	1.70	1.17	0.13				
7	CON	2.19	1.30	0.14				

Despite the small differences between grades on the total checklist score, results from the discriminant analysis indicated that checklist scores predicted grade. Similarly, the checklist subscores of Lexical Cohesion and Conjunction also predicted grade. Reference, however, did not (see Table 14). Scores on the checklist did not predict any special learning designation. This is likely due to the small numbers represented in these categories.

Table 14

Tests of Equality of Group Means for Checklist Scores

Subtest	Wilks' Lambda	F	df_1	df_2	p
REF	0.991	0.917	3	306	.433
CON	0.959	4.411	3	306	.005
LEX	0.911	9.992	3	306	.000
Total Score	0.938	6.698	3	306	.000

Concurrent criterion-related validity. I used the measures of writing fluency (WSC and TWW) and syntactical complexity (MLTU and SI) as indicators of concurrent criterion-related validity. These measures were calculated and recorded for each writing sample. I took two steps in determining the adequacy of these measures for detecting improvements in writing with grade within this data set. First, I calculated the means and standard deviations for each measure by grade. The results are presented in Table 15. The data show that measures of SI, MLTU, WSC, and TWW increased with grade.

Table 15

Means and Standard Deviations of TWW, WSC, MLTU and SI by Grade

Writing	Grade 4		Grade 5		Grade 6		Grade 7	
Measure	M	(SD)	M	(SD)	M	(SD)	M	(SD)
WSC	34.24	(13.46)	41.16	(14.53)	45.80	(15.47)	57.62	(14.47)
TWW	37.00	(13.33)	43.54	(14.49)	48.83	(15.22)	59.58	(13.76)
MLTU	7.88	(2.52)	7.94	(2.38)	8.87	(3.03)	9.53	(3.06)
SI	1.18	(0.26)	1.19	(0.27)	1.25	(0.38)	1.23	(0.21)

Note. M = mean; SD = Standard deviation

Next, I used the results of a discriminant analysis to determine if scores of WSC, TWW, MLTU, and SI were able to predict grade. A Wilks' Lambda test of equality of group means was conducted with all four scores. The results are presented in Table 16.

Table 16

Tests of Equality of Group Means for Writing Measures and Grade

Writing	Wilks'	F	df_1	df_2	p
Measure	Lambda				
TWW	.733	37.195	3	306	.000
WSC	.729	38.007	3	306	.000
MLTU	.942	5.976	3	293	.001
SI	.989	1.048	3	293	.372

Note. WSC = words spelled correctly; TWW = total words written; MLTU = mean length of T-Unit; SI = subordination index.

Results showed that MLTU, TWW and WSC were able to predict grade membership, while SI was not.

I then calculated correlations between the scores on the checklist and writing fluency measures of TWW and WSC and syntactic complexity measures of MLTU and SI using Pearson product moment correlations. As can be seen by examining Table 17, the relationships between fluency measures and the total checklist scores, Conjunction subscores and Lexical Cohesion subscores were significant (p < .001) These relationships had medium effect sizes. The relationships between MLTU and Total Test scores and Reference subtest scores were also significant (p < .05). Though these relationships were small, according to Cohen (1992) correlations between .1 and .3 denote a small but not trivial effect. SI showed no significant correlation to any of the subtest scores.

Table 17

Correlations Between Checklist Scores and Concurrent Measures

Score	WSC	TWW	MLTU	SI
REF	-.014 (.808)	-.005 (.929)	.203 (.000)	.041 (.477)
CON	.390 (.000)	.393 (.000)	-.065 (.259)	.034 (.558)
LEX	.335 (.000)	.336 (.000)	.173 (.003)	.040 (.487)
Total	.338 (.000)	.346 (.000)	.146 (.012)	.061 (.296)

Note. Total = total checklist score.

p value in parentheses.

Summary

This chapter described the results of the steps taken in developing and evaluating an instrument for assessing cohesion in the writing of children. In the preliminary developmental stage of this study, findings from qualitative analyses of the checklist were combined with data demonstrating the statistical performance of items and a pilot interrater study. I used these findings to make changes to the checklist before attempting a large sample field test of the instrument.

In the second portion of this study, the large scale sample evaluation of the checklist, I performed three more item analyses to examine the statistical performance of Checklist 2.0 and 2.1 versions, as well as to examine the relationship of items to subtests. I also gathered data reflecting the reliability of the instrument. These included measures of internal consistency and interrater reliability and agreement. Additionally, I gathered data to contribute evidence for the validity of the instrument. These included a factor analysis, a discriminant analysis, and a concurrent criterion-related validity study. Evaluation of and implications for these results are discussed in Chapter Five.

CHAPTER FIVE: DISCUSSION

The process used in this study for the development of an instrument to measure cohesion followed that which was outlined by Crocker and Algina (1986) and reported in Chapter Three. The remainder of this paper focuses on discussion of the outcomes of this process. The first topic to be discussed is that of the changes that were made to the checklist through the course of this investigation. The next area to be addressed will be the interpretation of the results of the reliability and validity studies. The validity section will include a discussion and interpretation of the concurrent criterion-related findings, the interpretation of subtest scores as it pertains to construct validity, and the discrimination ability of the checklist's scores. This chapter will conclude with a discussion of the limitations of this study and implications for future research and practice. The section discussing the implications for future research contains recommendations for further modifications to the content and form of Checklist 2.1 for prospective development of the instrument. This section also includes some suggestions for future reliability and validity studies.

Summary of the Checklist Item Development

The preliminary form of this checklist (Checklist 1.0) consisted of 27 items grouped under the four subheadings of Reference, Conjunction, Lexical Cohesion, and Global Cohesion. Through the process of the preliminary development and large scale evaluation, the checklist underwent several revisions with the final form consisting of only 13 items and 3 subtests (Checklist 2.1). The changes made consisted of eliminating the subsection on Global Cohesion, combining items in the Conjunction subsection, and deleting and combining Lexical Cohesion items.

Global Cohesion was no longer part of Checklist 2.1 even though aspects of global cohesion play a role in developing a coherent and cohesive piece of writing (McCutchen &

Perfetti, 1982; Smith, 1999). Several factors contributed to the decision to eliminate this subsection of the test. The results of the item analysis indicated that almost all writing samples in this study received credit on the items for sequential organization and consistent use of tenses. Conversely, very few writing samples received credit for the use of paragraphs. Given the length of the writing samples and the timed nature in which they were administered, the lack of paragraph structures was not surprising. Many of the samples were not long enough to warrant more than one paragraph. For those that were, with only one sheet of paper on which to write and three minutes to complete the writing task, time and space constrictions may have played a role in reducing the tendency to use paragraphs. Similarly, the limited length of the writing samples may have reduced the likelihood of errors with consistent tense use.

Due to their high or low proportions of credit, the items regarding sequential organization, consistency of tenses and use of paragraphs were considered too easy or too difficult. Items that are too easy or too difficult are not effective in discriminating among students (Sax, 1997). As most writing samples scored the same on these items, the information was not felt to be helpful in determining differences between the abilities of individual students to use cohesive devices. Subsequently, they were removed from the checklist. The remaining item regarding implied causal relationships was eliminated despite its adequate performance on the item analysis as it did not seem logical to retain one item in a subtest.

These Global Cohesion items may have been more important if different or longer writing samples had been used or if a different genre of writing was attempted. McCutchen and Perfetti (1982) indicate that even the most immature writers have knowledge of the narrative form and that knowledge of the text form contributes to the overall coherence achieved in writing. Perrera (1984) concurs that when children begin to write coherent texts, they have more success with chronologically ordered texts. As these overall organizational patterns have not been found to present problems in children's narrative writing, it is not surprising that the items from this portion of the checklist did not reveal many differences between writers. It may be that these items reflecting global cohesion would be more useful in discriminating between writers if the text form was not as familiar. This aspect of cohesion may be more informative when examining expository, persuasive or descriptive pieces of writing.

The Conjunction subsection was another area that was changed substantially through the course of this study. Interestingly, the results of the item analysis supported the findings of other researchers in regards to the common occurrence of the conjunctions 'and,' 'then,' and 'so' in children's writing. These were clearly the most frequently occurring conjunctions as indicated by the proportion of writing samples receiving credit for these items. In fact, most other forms of temporal, additive, and causal conjunctions occurred rarely. Consequently, I made the decision to collapse items for individual conjunctions into larger category groupings. These combinations resulted in improvements in the discrimination of most items.

Two items, however, remained problematic. These items reflect the use of subordinating conjunctions used to show temporal connections between sentences and clauses. The poor discrimination indices on these items, however, is likely due to the relatively low proportion of use of such conjunctions by the writers in this sample. These items were still considered valuable to the checklist as their deletion would result in a loss of information regarding the use of temporal conjunctions which have been noted to be common in the writing of children (Crowhurst, 1987; Perrera, 1984; Smith, 1999). Furthermore, removal of these items would result in a gap in the kinds of conjunctive cohesion measured by this instrument. These items may be more valuable in evaluating cohesion in more typical curricular writing as well as in different writing genres. Perrera, for instance, found that children did not use subordinating conjunctions as much in story writing as they did in other kinds of writing tasks.

The final area of modification occurred in the subsection of Lexical Cohesion. Converses and antonyms rarely appeared in the writing samples so the item reflecting these types of cohesion was deleted. Superordinates, synonyms and near-synonyms also appeared relatively infrequently. However, it was found that by combining the two items regarding this form of cohesion, the discrimination index was improved with limited loss of information. The item now reflected whether or not a student was using this form of cohesion, but did not reflect the 'closeness' of the ties. As the writing samples used here were quite short, the distance between ties was rarely great. As physical proximity impacts the degree of cohesive bond formed (Halliday & Hasan, 1976), in a longer sample, this combined item might be problematic.

All other items on the checklist functioned in an acceptable manner as judged from the item analysis and subsequently were left unchanged in Checklist 2.1. It should be noted however that two problems became apparent during the scoring process. These problems primarily affected the scoring of the Reference subsection. One problem involved item 5 which reflected the use of cohesive devices on a sentence by sentence basis. In most cases, these items were not problematic. However the scoring of this item was impacted by shifts in the story. It was the observation of this researcher that when there was a shift in story events, an anaphoric reference between the last sentence of one segment, and the first sentence of the next one was not warranted. Therefore, writing samples that contained shifts in time, place, or speaker may have been inadvertently penalized on this item.

The other problem related to the use of stories told in first person voice. As this instrument focused on endophoric forms of reference (reference within the text) and the referent for 'I' is exophoric (external to the text) repetitions of the first person pronoun 'I' were not given credit as forms of reference. Stories containing only first person pronouns would not receive credit on the first two checklist items and would consequently receive lower scores. It was rare, however, to find writing samples that contained no examples of third person pronouns.

These problems do not affect the scoring of individual writing samples but may impact the use of the checklist for cross student comparisons. Caution should be taken when comparing texts that are written in first and third person voices or in comparing stories with dialogue or other shifts in events with stories that do not contain such elements.

Findings for Reliability and Validity

Interpretation of the findings of this study lead to many important considerations and conclusions beyond the statistical performance of checklist items. Evaluation of an assessment instrument does not end with the analysis of the items. Reliability and validity of an instrument also require examination. These areas are addressed in the next section.

Reliability

I assessed reliability by examining internal consistency and interrater reliability and agreement. The alpha levels provided an indication of the internal consistency of the instrument. Although an alpha of .32 is low, this may be a function of the checklist length. Sax (1997) indicates that reliability increases with the number of items on the test. Therefore a reliability measure of .32 given that there were only 13 items on the checklist may be considered reasonable. Similarly, the alpha levels for each of the subtests were not very high. Again, each subtest consisted of only a small number of items. The Conjunction subscore showed particularly low consistency among its items as reflected by its low alpha (.22). and large SEM (1.05). As mentioned earlier, two items on the Conjunction subtest were still performing poorly on the items analysis and this may have affected the internal consistency of this subtest. Lexical cohesion also demonstrated a low alpha but this figure is difficult to interpret given that there were only two items on this last section of the instrument. The issue of checklist length may have been further compounded by the limited length of the writing samples.

Several factors may have impacted the internal consistency of this instrument by reducing the variability of the scores. A reduction in the variability of the sample will result in a reduction in the reliability findings (Sax, 1997). One way variability in scoring is reduced is with increased item difficulty, as many writing samples will receive the same score on a difficult item (Sax, 1997). According to this item analysis results, 7 of the 13 checklist items are considered difficult. With so many items showing up as difficult, the overall variability of scores may have been compromised. Furthermore, an examination of the differences in overall test and subtest scores revealed that mean scores did not vary greatly across grades. Similarly, the small number of items on each subtest further reduced the scoring variability.

The internal consistency of the 13 item Checklist 2.1, as found in this study, was weak. Reliability could be improved by including more easy items and increasing the number of items overall. Increasing the number of items on which the checklist is scored could also be accomplished by scoring more than one writing sample and pooling the results rather than increasing the number of checklist items. Furthermore, as checklist scores reflected the samples that were used to test the instrument, internal consistency scores may be different with different writing samples.

Further evidence regarding the checklist's reliability was provided through examination of interrater reliability and agreement. Review of the literature did not provide absolute guidelines as to how much interrater reliability or agreement is considered adequate. I considered the overall levels of interrater reliability attained for Checklist 2.1 adequate due to the lack of significant differences among raters as determined by the one-way ANOVA and the correlations between rater pairs. However, it is important to note that the process of training raters for this interrater study was paramount in establishing agreement.

The amount of interrater reliability and agreement required depends on the uses of the instrument. Higher levels than were attained here may be desired if an examiner wished to compare checklist scores to those obtained by another examiner.

In addition to providing information about the reliability of the measure, checks of interrater agreement also provided some insight as to which checklist items could still be considered ambiguous. Two items (item 5 and item 13) demonstrated noticeably lower levels of agreement than others and may warrant some further editing for the purposes of clarification.

Validity

Studies of an instrument's validity provide evidence that the instrument is measuring the construct it claims to measure. This may be accomplished in several ways. Evidence for criterion-related validity demonstrates that the scores on an instrument correlate with some related external criteria. Evidence for discriminant validity demonstrates the difference between what the test measures and different constructs. A factor analysis can also support construct validity by providing evidence for the relationships between items that reflect a single construct. Demonstration that the construct is important and can be measured also support arguments for validity.

Concurrent criterion-related validity. Prior to conducting the validity study, I examined the performance of the measures of TWW, WSC, MLTU and SI to determine their adequacy as measures of writing proficiency in this data set. Evidence showing the growth and discrimination ability of TWW, WSC and MLTU scores suggested that these measures reflected growth in writing ability across grades. This provided further evidence beyond what was reported in the literature review as to the relationship of these scores to developmental growth in writing proficiency. I then correlated these scores with scores from the checklist in an attempt to provide one source of evidence for concurrent criterion-related validity.

Cohen (1992) states that correlations above .1 and below .3 demonstrate a small but non-trivial effect. Correlations between .3 and .5 demonstrate medium effect sizes. These descriptors apply to the practical significance of a correlation value. Given Cohen's definition of practical significance of correlations, the two scores of TWW and WSC were found to show medium effect size correlations to the Total Test score on the checklist while MLTU showed a small size correlation to the Total Test score. No relationship was found with SI.

The failure of SI to show relationships to any of the other writing measures resulted primarily from the small variation in these scores across grades. The difference between Grades 4 and 7 on this measure was only .05 clauses per T-unit. Similarly this measure did not predict grade membership. SI growth as measured in previous studies has been shown to increase across grades but with some fluctuations in the growth pattern (Scott, 1988) as was also found here. While Hunt (1965) found a more noticeable pattern of growth in SI, she was looking at 4 years grade difference between groupings of students (Grades 4, 8 and 12). Even Klecan-Aker and Lopez's (1985) study of SI differences between students in grades with 3 years apart (Grades 6 and 9) found no statistical difference between the scores of the two groups. This lack of variability between grades would give this measure very little discriminating power, and make it difficult or impossible to detect any relationship it may have to other growth measures.

Similarly, limited variability in scores of MLTU may also have impacted the size of their relationship to cohesion scores. MLTU and SI have generally been calculated from much larger writing samples. Hunt (1965), for example, used writing samples 1000 words long to calculate MLTU and SI. Although the MLTU showed growth with increased grade and SI showed a small growth pattern as well, it is possible that these values would have been more precise had the writing samples from which they were calculated been longer.

While the Total Test score showed relationships to other measures of writing, these values were not large and therefore are not strong indicators of concurrent criterion-related validity. However, relationships between MLTU, TWW and WSC, and the cohesion score may provide evidence for another kind of validity.

Discriminant validity. Discriminant validity is indicated by evidence showing how the construct in question differs from other constructs. The size of the correlations between the scores on the cohesion checklist and measures of writing fluency and syntactic complexity provided evidence of discriminant validity.

It was argued earlier that measures of cohesion should show some relationship to measures of writing fluency and syntactic complexity as they all reflect skills related to writing proficiency. But while all measures may reflect writing ability, they each reflect different aspects of that ability.

Several factors could explain diverse performance on different measures of writing. These differences relate to the underlying skills involved in various aspects of writing. For instance, while researchers have called attention to the difficulties students with language-learning disabilities have with cohesion, syntax and other general areas of writing, not all children with writing difficulties have an underlying language problem. Children with nonverbal learning disabilities and those with dysgraphia may have problems that are associated with the motor aspects of writing (Richards, 1999; Thompson, 1997). These children may demonstrate difficulties with writing speed and letter formation which may impair scores of writing fluency, but may not impact the ability to write coherently or in complex sentence forms. Additionally, it has been my professional observation that children having language impairments that primarily impact the pragmatic aspects of language may have significant difficulty with cohesion but not have difficulty with writing fluency or written syntax. In each of these scenarios, cohesion scores, syntax scores and writing fluency scores would not be closely related.

The modest size of the relationships found between TWW, WSC, MLTU and the Total checklist scores provided evidence of discriminant validity. That is, the cohesion checklist did not measure the same skills as writing fluency or syntax measures. If it did, we would expect higher correlations. The medium and small effect sizes found here suggest that cohesion scores are related to measures of writing fluency and syntactic complexity, as all three measures relate to writing proficiency, but are not measures of the same underlying construct. In fact, it was this opinion that prompted this study. If mechanical skills in writing were highly reflective of discourse level skills such as the use of cohesion devices, there would be no need to measure cohesion separately, as measurements of mechanical skills would be ready made indicators of cohesion. These results show that cohesion is a separate writing skill that can be measured.

Discrimination ability of checklist scores. The ability of an assessment tool to detect differences between students of differing abilities provides further evidence for validity. The total checklist score was able to predict grade membership. However, it was not able to predict the special learning designations of ESL/D, SLR, LD, and Other, either individually or combined. One explanation for this lack of predictability could be the small numbers of writing samples generated by children with special learning designations in this data source. Of the 312 writing samples used, there were only 7 samples generated from children designated SLR, 30 samples from children designated ESL/D, 21 samples from children designated LD, and only 9 from children designated as Other. With such small numbers represented in these groups, it could not be established whether performance on this instrument detected differences between these students and their normally achieving peers. Sampling across four grades also made this kind of detection difficult. For example, there may be minimal differences between a high performing student in Grade 4 and a student in Grade 7 with a special learning designation. I did not perform discriminant analyses by grade as the number of writing samples generated by children with special learning designations in each grade was too small.

Although the Total Test scores were able to predict grade membership, differences existed among subtests in their ability to do so. For instance, the Reference subscore could not be shown to predict grade membership at all. One explanation for this could be that the scores on this section of the checklist were not sensitive to incremental developmental growth as would be expected from grade to grade. Perrera (1984) has suggested that "pronominal reference is used early and extensively in children's writing" (p. 241). If this is true, then items of reference may have difficulty predicting grade differences in the upper elementary years simply because children have already developed their use of this form of reference. However, Perrera also stated that children have ongoing difficulties with pronominal agreement and using pronouns in an unambiguous way but she did not indicate at what developmental stages these errors diminish.

Another explanation for this lack of predictability could be related to the problem encountered when scoring Reference items across shifts in story events. It is possible that some older students used Reference items with less error and ambiguity but lost credit due to dialogue use or shifts in story events. This could result in older and younger writers with similar Reference scores from credit on different items. This explanation may partially account for the lack of substantial differences among grades on the mean Reference scores.

Another explanation for the inability of reference scores to predict grade may reflect the "all" or "none" scoring criteria applied in this section. Older students with only one error on an item would receive the same item score as younger students with multiple errors, yet it may be argued that these two children may have differing levels of ability in this area.

While the Reference subscore showed poor discrimination among grades, the Conjunction and Lexical Cohesion subscores did not. Their ability to predict grade membership suggested that there is some relationship between performance on these two subtests and developmental writing ability.

Subscores versus total scores. Evidence for construct validity also comes from findings that support the underlying theorized construct in question, in this case the components of cohesion reflected in the subtests of the instrument. The subsections of the checklist were based on the concepts outlined by Halliday and Hasan (1976). Results of the item analyses showed a difference in the performance of the items as parts of the total checklist when compared to the item performance as parts of the subtests. This finding seems to support the relationship between the checklist subsections and the underlying constructs of referential, conjunctive, and lexical cohesion described by Halliday and Hasan. This relationship is evidenced not only by the improvement in items scores when analyzed as part of the subtests, but also by the lack of relationship between the three subsets of items.

Further evidence regarding the differences in subtest performance is indicated by variation in how the subtests related to other measures of writing. While the checklist Total Test score showed correlations to three other writing measures, individual subtests of the instrument varied in their relationship to these same three measures. For instance, the Reference subscore was shown to have a small effect size correlation to MLTU. Conversely, it showed no relationship to fluency writing measures while the Lexical Cohesion and Conjunction subscores did. This suggests to me that the length of a writing sample has no influence on whether or not a child successfully or unsuccessfully used devices of referential cohesion. On the other hand, the relationship between writing fluency measures and Conjunction and Lexical Cohesion may be a function of the length of the writing samples. It may be that the longer the writing samples were, the more variety there was in the vocabulary and the kinds of conjunctions used. Another explanation is that Conjunction and Lexical subscores are, in fact, generally related to overall writing proficiency as are TWW and WSC.

The Conjunction subscore was the only one that did not correlate to MLTU. This may have resulted from higher uses of coordinating conjunctions. The use of more varied coordinating conjunctions and less subordinating ones could result in high cohesion scores coupled with lower MLTU as the MLTU generally increases with increased subordination.

The Lexical Cohesion subscore was the only one to show correlations to all the three measures of TWW, WSC and MLTU. These findings, though mixed, do provide evidence supporting the argument that facility with cohesion is related to proficiency with several different underlying skills related to three areas of cohesion described by Halliday and Hasan (1976).

Because of the findings regarding the differential performance of checklist subtests, I expected that a factor analysis would provide a solution of three components underlying the checklist items. However, a factor analysis was unable to account for a large portion of the variance in checklist scores and did not show substantial variable loadings on each component. Loadings greater than .30 in absolute value are generally considered significant (Academic Computing And Instructional Technology Services, 1995) but Stevens (1996) indicates that components require a minimum of four loadings greater than .60 or a minimum of three loadings greater than .80 to be reliable. There should be a minimum of three observed variables for each factor and, ideally, each variable should load significantly on a single factor (Academic Computing And Instructional Technology Services, 1995).

One explanation for the failure of the factor analysis may relate to the small inter-correlations between checklist items. The reason for the poor correlations found between checklist items may lie in the dichotomous nature of the variables used on this instrument. Gorsuch (1983) explains:

> When data are noncontinuous, it is possible for several individuals to receive exactly the same score on one variable. However, if these same individuals do not receive the same score on another variable, the two variables cannot correlate perfectly even if the underlying relationship is perfect. The reduction in correlation occurs most often with dichotomous variables because a great number of individuals receive the same score (pp. 291 - 292).

Another explanation for the difficulty with interpreting the factor analysis relates to the quality of the data. Error in the data can strongly influence the results of a factor analysis, therefore, the instrument used for a factor analysis needs to be reliable (Academic Computing And Instructional Technology Services, 1995; Tabachnick and Fidell, 1996). As the internal consistency results on this instrument were not strong, this may have impacted the outcome of this analysis.

Despite the difficulty with interpretation of the factor analysis, the findings of the item analysis seem to support the subtest divisions of the checklist. The ITEMAN analysis is suited to dichotomous variables. Additionally, point biserial correlations are based on correlations between a dichotomous variable (item score) and a continuous variable (subtest or total score). Given the improved point biserial correlations with the checklist subscores and the differential performance of subtests, interpretation of checklist scores may be better served by examining performance on each of the subsections individually. These findings may also suggest that cohesion is not a single construct as was first expected here, but may be made up of several unrelated or semi-related latent skills or abilities. Consequently, subtest scores may need to be considered separately.

Contributions of this Research

This study formed the initial stages of developing a checklist to use in evaluating cohesion in writing. Through this research, the items on the checklist have been revised to reduce ambiguities and improve their performance on a classical item analysis. The final interrater study showed adequate agreement among raters. It has been shown that the checklist's total score is able to predict grade membership thus showing its sensitivity to differences in the writing of children of different grade levels. Additionally, checklist subscores of Conjunction and Lexical Cohesion were able to predict grade membership. As well, checklist scores demonstrated discriminant validity in their relationships to other measures of writing proficiency. There is also evidence to suggest that subtests be scored independently.

While further development of the checklist is still warranted, this research has contributed to the field in three main ways. First, the study done here provided the ground work for further development of an instrument to measure cohesion in writing. Second, information on writing development and evaluation is sparse in the literature. This study will add to that body of knowledge through its examination of evaluating cohesion in writing. Third, the number of studies on cohesion in the writing of school-aged children is also limited. This study of cohesion contributes to that body of knowledge.

Limitations

As this study warranted comparison of the performance of checklist items across many writing samples, a single type of writing sample was used. I felt that the type of writing samples chosen should be as homogenous as possible to make comparisons between writing sample scores more clear. I chose CBM samples as they met this criteria for homogeneity and were available in large quantities across grades. However, CBM writing samples are short and administered under time constraints. No proof-reading or editing is allowed. The performance of items on this testing of the checklist was indeed limited by the constraints under which these writing samples were generated. It is expected that item analyses conducted with untimed edited narrative writing samples would have different results.

Another necessary limitation of the sample chosen was the genre used. Many studies have shown that the types of cohesive devices used in writing are related to the genre of the written text (Crowhurst, 1981, 1987; Hidi & Hildyard, 1983; McCutchen & Perfetti, 1982; Pellegrini et al., 1984). A single genre was used as it would be difficult to interpret an item analysis based on comparing different kinds of writing samples. That is, because different devices are genre specific, variability in cohesion scores found with mixed writing samples may have reflected differences not related to proficiency with the use of cohesive devices. By using only one genre, analysis of checklist items was made easier, but items that may have been more important in detecting differences in other writing genres were lost. The checklist developed here, consequently, may only be useful in evaluating cohesion in narrative writing samples.

The sample of students who generated the writing also impacted the outcomes of this study. As the samples were generated from only three schools, there were only 67 samples generated from children designated with special learning needs or learning disabilities. With such small numbers represented in this group, it could not be established whether performance on this instrument detected differences between these students and their normally achieving peers. Sampling across four grades also made this kind of detection difficult. Consequently, there is not enough information to determine at this point whether such an instrument would be useful in detecting differences between different groups of students in the same grade. This ability would be crucial for its value as a diagnostic tool; that is, in its ability to show differences between a target student and same grade peers, and to detect growth in a single student over short periods of time.

Implications for Future Research

The findings here reflect the first stages in the development of a checklist for evaluating cohesion in writing. Further development of the instrument is warranted before its value as an educational tool can be determined. In addition to the suggestions made throughout this discussion, some further suggestions for future research are explored here.

Proposed Changes to Checklist Content and Format

Content. There are still some items on the checklist that may benefit from further modifications. These include items that showed poor interrater agreement and items on the Reference subsection which were presenting problems when scoring stories with shifts in events. For instance, Item 5 could be reworded to say "Except in topic sentences, each sentence is connected to the one proceeding it by at least one form of reference." Improvements could also be made to the Reference section by setting criteria rather than "all" for credit on an item. For example, criteria for credit could include an allowable number or proportion of violations per designated number of T-units. Establishing appropriate ratios would require testing samples of writing with this subtest and determining which proportions reflected the best discrimination between groups of learners.

Item 13, which addressed the use of complementary lexical items, was also presenting difficulties with rater agreement. The scoring guide could include more explicit scoring instructions for this item such as by including a systematic way of detecting examples of collocation. This could include a procedure like underlining all the nouns and verbs and examining them to find word pairs that meet the definition of collocation.

As the internal consistency of the checklist may be better with more items, items that had been combined in order improve this item analysis using the very short three- minute narrative writing samples may be separated into a greater number of discreet items to be tested with longer samples or writing of other genres. In particular, the Conjunction items may be separated into more discreet items and the subtest of Global Cohesion could be reintroduced.

Format. As the length of the sample seemed to impact scores on certain subtests more than others, some guideline reflecting the length of the sample to be evaluated may be prudent. This guideline could form an minimum requirement for length. Additionally, where an examiner wished to use the checklist to evaluate longer samples of writing, only a portion reflecting the length requirement need be scored. For example, the examiner could score the first 50 T-units. This would not only help to control for differences in scores caused by length, but also would make the task of scoring more manageable. Another way to control for size would be to score Conjunction and Lexical Cohesion subtests on the basis of a proportion rather than an absolute score.

Proposed Procedures for Checklist Evaluation

As the genre, degree of editing, and audience all affect the types of cohesive devices used it is recommended that the checklist be evaluated with a variety of writing samples including more typical curricular narrative samples. Furthermore, by testing the checklist with other forms of writing samples, it could be established how much of the validity and reliability problems encountered in this study were related to the checklist and how much could be accounted for by the writing samples used in this study. I have included some suggestions for future analysis of the checklist's performance.

One suggestion is to use writing samples from the Foundations Skills Assessment administered provincially to Grades 4 and 7 . The advantage of this choice is that these samples are administered in a standardized way, and their scores could be used to establish concurrent criterion-related validity. Furthermore, by examining the checklist performance within large numbers in each grade, it may be possible to establish whether or not the checklist can detect differences between same grade peers of differing ability. One limitation of this choice is that it would not be possible to examine performance on the checklist across grades.

Another suggestion for further evaluation of this instrument is the District Writing 5 Exam. This exam is administered annually to all Grade 5 students in School District # 57. Each sample is rated on a 4 point holistic rating scale. The advantage of this choice is that these samples are administered in a standardized way. As each sample is only rated on a four point scale, these scores are not diverse enough to serve as criteria for concurrent validity. However, a discriminant analysis could determine whether checklist scores would be predictors of the holistic rating. This would provide evidence for the checklist's ability to detect differences between same-grade peers which is paramount for its use as a diagnostic device. One limitation of this choice is that results could not be generalized beyond Grade 5.

Another source of writing samples on which to test the instrument is writing portfolios. The challenge with this choice would be selecting samples that are comparable on the basis of genre, degree of editing, instruction, and audience. The advantage of this selection would be the authenticity of samples and the opportunity to determine how the checklist detects cohesion in "best" samples of writing. The instrument also could be used to explore cohesion in "draft" and "published" versions of the same writing included in the portfolios, thus providing pedagogical guidance. Use of portfolios would also allow for testing across grades, as well as providing an indicator of the writing development of individual students over the school year. The use of such portfolio-based writing samples would provide an important indicator of practical or clinical validity.

Reliability. It would be valuable to determine if longer writing samples, or those produced without time constraints, resulted in larger values for internal consistency. Furthermore, evidence of the stability of cohesion scores across time through test-retest procedures and equivalence of cohesion scores across writing samples of similar genre and instructional approach would be valuable in the development of this tool.

Validity. Further assessment of the validity of this instrument is also warranted. In order for this tool to be used diagnostically, for instance, there needs to be evidence showing the checklist's ability to predict special learning designations. This may be best accomplished by using a disproportional stratified random sample that represented large proportions, and therefore large samples of children with these designations. This type of sampling is useful when comparisons among groups is of interest (Palys, 1997). An important consideration would be to use widely accepted criteria for the identification of specific designations of special needs.

Another area to be addressed is concurrent criterion-related validity. The argument for the concurrent validity of this instrument may be strengthened by a comparison of scores from the cohesion checklist to holistic ratings of readability of the same samples, as this aspect of writing is expected to be more related to cohesion (Lindeberg, 1984; Rutter & Raban, 1982; Zarnowski, 1981) than the measures used here. This may include procedures such as comparing the checklist scores to teachers' ratings of quality or to the analytic scoring rubrics used in the Writing Reference Set (British Columbia Ministry of Education, 1996b).

Once it could be demonstrated that the checklist was able to detect differences between writers of the same age with differing abilities, validity for use of this tool as a diagnostic instrument that can be used to establish and monitor progress of intervention goals still would need to be determined. This would involve pre- and post- treatment measures to determine if the checklist was sensitive to changes in the use of cohesion over time.

Implications for Practice

The initial impetus for this research was to create an instrument that could be used by speech-language pathologists to detect and define problems with cohesion in authentic writing samples. That is, the instrument, in its completed form, would assist in first detecting which children were having difficulty in using cohesive devices in writing when compared directly to their peers on the same writing task. Second, I wanted an instrument that could reveal which aspects of cohesion were lacking or problematic in a child's writing. Third, I wanted the instrument to be able to detect differences in an individual's use of cohesive devices with intervention.

Although the checklist developed here is not yet ready for these uses, it still may have application as a reference guide for observing children's writing. In this way it could assist an examiner in describing what types of cohesion the child is using. In addition to this use, the results of this study can inform professionals who work with children on their writing skills in two main ways. First, the results of this study suggest that different kinds of cohesion may benefit from separate evaluation. Proficiency in the use of referential cohesion may develop quite differently and reflect a different type of skill than that seen with the use of lexical devices or conjunctions. Second, writing is a complex process requiring facility with a number of different skills. Assessment in any single area will not tell us much about a writer's overall writing ability. Writing ability constitutes more than a single latent variable, therefore assessment across a variety of skills is necessary to get an adequate picture of a writer's abilities, disabilities, strengths, and weaknesses. Referential cohesion, conjunction, and lexical cohesion are only small parts of a complex process or skill.

REFERENCES

Academic Computing and Instructional Technology Services. (1995). Factor analysis using SAS PROC FACTOR [On-line]. Available: http://www.utexas.edu/cc/docs/stat53.html

Anderson, P. L. (1982). A preliminary study of syntax in the written expression of learning disabled children. Journal of Learning Disabilities, 15(6), 359-362.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education, 5(1), 7-73.

British Columbia Ministry of Education (1996a). English language arts K to 7 integrated resource package. Victoria, BC: Author.

British Columbia Ministry of Education (1996b). Evaluating writing across the curriculum: Using the writing reference set to support learning. Victoria, BC: Author.

Calfee, R. C. & Freedman, S. W. (1996). Classroom writing portfolios: Old, new, borrowed, blue. In R. C. Calfee & P. Perfumo (Eds.). Writing portfolios in the classroom: Policy and practice, promise and peril. (pp. 3-26). Mahwah, NJ: Lawrence Erlbaum Associates.

Canter, A. & Marston, D. (1998). Helping children at home and school: Handouts from your school psychologist. Bethesda, MD: National Association of School Psychologists.

Carrow-Woolfolk, E. (1996). Oral and Written Language Scales. Circle Pines, MN: American Guidance Service.

Choate, J. S. & Miller, L. J. (1992). Curricular assessment and programming. In J. S. Choate, B. E. Enright, L. J. Miller, J. A. Poteet, & T. A. Rakes (Eds). Curriculum-based assessment and programming (2nd ed) (pp. 43-77). Needham Heights, MA: Allyn and Bacon.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112 (1), 155-159.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart & Winston.

Crowhurst, M. (1981) Cohesion in argumentative prose written by sixth-, tenth-, and twelfth-graders. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles. (ERIC Document Reproduction Service No. ED 202 023)

Crowhurst, M. (1987). Cohesion in argument and narration at three grade levels. Research in the Teaching of English, 21 (2), 185-197.

Dagenais, D. J. & Beadle, K. R. (1984). Written language: When and where to begin. Topics in Language Disorders, 4 (2), 59-85.

Deno, S. L., Marston, D., & Mirkin P. (1982). Valid measurement procedures for continuous evaluation for written expression. Exceptional Children, 48 (4), 368-371.

Educational Testing Service (1993). The ETS collection catalog :Achievement tests and measurement devices (2nd Ed.). Phoenix, AZ: Oryx Press.

Engelhard, G. (1998). Review of the CTB Writing Assessment System. In J. C. Impara & B. S. Plake (Eds.) The thirteenth mental measurements yearbook (pp. 329-331). Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska.

Englert, C. S. & Raphael, T. E. (1988). Constructing well-formed prose: Process, structure and metacognitive knowledge. Exceptional Children, 54 (6), 513-520.

Fewster, S. (2000). School based evidence for validity of curriculum-based measurement norms. Unpublished master's thesis, University of Northern British Columbia, Prince George, British Columbia, Canada.

Gillam, R. & McFadden, T. U. (1994). Redefining assessment as a holistic discovery process. Journal of Childhood Communication Disorders, 16 (1), 36-40.

Gorsuch, R. L. (1983). Factor Analysis. (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Graham, S., Harris, K. R., MacArthur, C. Schwartz, S. (1998). Writing instruction. In B. Y. L. Wong (Ed.) Learning about learning disabilities (2nd ed., pp. 391-423). San Diego: Academic Press.

Greenberg, K. L. (1987). Defining, teaching and testing basic writing competence. Topics in Language Disorders, 7 (4), 31-41.

Halliday, M. A. K. & Hasan, R. (1976). Cohesion in English. London: Longman Group.

Hansen, J. B. (1998). Review of the Test of Written Language - Third Edition. In J. C. Impara & B. S. Plake (Eds.) The thirteenth mental measurements yearbook (pp. 1070 - 1072). Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska.

Hedberg, N. L. & Fink, R. J. (1996). Cohesive harmony in the written stories of elementary children. Reading and Writing: An Interdisciplinary Journal, 8, 73-86.

Hidi, S. E. & Hildyard, A. (1983). The comparison of oral and written productions in two discourse types. Discourse Processes, 6, 91-105.

Howell, K. W., Fox, S. L. & Morehead, M. K. (1993). Curriculum-based evaluation: Teaching and decision making, (2nd ed.). Belmont, CA: Wadsworth.

Hughes, D., McGillivray, L. & Schmidek, M. (1997). Guide to narrative language: Procedures for assessment. Eau Claire, WI: Thinking Publications.

Hunt, K. W. (1965). Grammatical structures written at three grade levels. Champaign, IL: National Council of Teachers of English.

Impara, J. C. & Murphy, L. L. (Eds.). (1994). Psychological assessment in schools. Lincoln, NE: University of Nebraska Press.

Impara, J. C. & Plake, B. S. (Eds.). (1998). The thirteenth mental measurements yearbook, Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska.

Isaacson, S. (1991). Assessing written language skills. In C. S. Simon (Ed.) Communication skills and classroom success: Assessment and therapy methodologies for language learning disabled students, (pp. 224-237). Eau Claire, WI: Thinking Publications.

ITEMAN (Version 3.50) [Computer software]. (1994). St. Paul, MN: Assessment Systems Corporation.

Kimmel, E. W. (1998). Review of the Writing Process Test. In J.C. Impara & B.S. Plake (Eds.) The thirteenth mental measurements yearbook (pp. 1160-1161). Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska.

King-Sears, M. E. (1994). Curriculum-based assessment in special education. San Diego: Singular Publishing Group.

Klecan-Aker, J. S. & Hendrick, D. L. (1985). A study of the syntactic language skills of normal school-age children. Language, Speech, and Hearing Services in Schools, 16 (3), 187-198.

Klecan-Aker, J. S. & Lopez, B. (1985). A comparison of T-units and cohesive ties used by first and third grade children. Language and Speech, 28 (3), 307-315.

Liles, B. Z. (1985). Cohesion in the narratives of normal and language-disordered children. Journal of Speech and Hearing Research, 28, 123-133.

Liles, B. Z., Duffy, R. J., Merritt, D. D. & Purcell, S. L. (1995). Measurement of narrative discourse ability in children with language disorders. Journal of Speech and Hearing Research, 38, 415-425.

Lindeberg, A. C. (1984). Cohesion and coherence in short expository essays. Proceedings from the Nordic Conference for English Studies.

Loban, W. (1976). Language development: Kindergarten through grade twelve. Urbana, IL: National Council of Teachers of English.

Marston, D. B. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. R. Shinn (Ed.) Curriculum based measurement: Assessing special children (pp. 18-78). New York: Guilford Press.

Marston, D. B. & Deno, S. (1981). The reliability of simple direct measures of written expression (Research Report # 50). Minneapolis: University of Minnesota Institute for Research of Learning Disabilities.

Miller, L. E. (1999). Evaluating English writing at the highschool and college level. Unpublished paper. (Available from Lois E. Miller, P.O. Box 101, Nass Camp, BC, V0J 3J0).

McCutchen, D. & Perfetti, C.A. (1982). Coherence and connectedness in the development of discourse production. Text, 2, 113-139.

Moss, P. A. (1995). Themes and variations in validity theory. Educational Measurement: Issues and Practice, 15 (3), 5-13.

Murphy, L. L., Impara, J. C. & Plake, B. S. (Eds.). (1999). Tests in print: An index to tests, test reviews, and the literature on specific tests. Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska.

Murray-Ward, M. (1998). Review of the Test of Written Expression. In J. C. Impara & B. S. Plake (Eds.) The thirteenth mental measurements yearbook (pp. 1067-1068). Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska.

Nelson, N. W. (1994). Curriculum-based language assessment and intervention across the grades. In G. P. Wallach & K. G. Butler (Eds.) Language learning disabilities in school-aged children and adolescents: Some principles and applications (pp. 104-131). Toronto: Maxwell Macmillan Canada.

Palys, T. (1997). Research decisions: Quantitative and qualitative perspectives. Toronto: Harcourt Brace.

Pellegrini, A. D., Galda, L. & Rubin, D. (1984). Context in text: The development of oral and written language in two genres. Child Development, 55, 1549-1555.

Perrara, K. (1984). Children's writing and reading. Oxford: Basil Blackwell.

Poplin, M., Gray, R., Larsen, S., Banikowski, A. & Mehring, T. (1980). A comparison of components of written-expression abilities in learning disabled and non learning disabled students at three grade levels. Learning Disabilities Quarterly, 3 (4), 46-53.

Poteet, J. A. (1992a). Educational assessment. In J. S. Choate, B. E. Enright, L. J. Miller, J. A. Poteet, & T. A. Rakes (Eds). Curriculum-based assessment and programming (2nd ed) (pp. 1-21). Needham Heights, MA: Allyn and Bacon.

Poteet, J. A. (1992b). Written Expression. In J. S. Choate, B. E. Enright, L. J. Miller, J. A. Poteet, & T. A. Rakes (Eds). Curriculum-based assessment and programming (2nd ed) (pp. 231-271). Needham Heights, MA: Allyn and Bacon.

Principles for fair student assessment practices for education in Canada. (1993). Edmonton, AB: Joint Advisory Committee. (Available from the Joint Advisory Committee, Centre for Research in Applied Measurement and Evaluation, 3-104 Education Building North, University of Alberta, Edmonton, AB, T5G 2G5).

Psychological Corporation (1992). Wechsler Individual Achievement Test. San Antonio, TX: Harcourt Brace and Company.

Ratner, V. L. & Harris, L. R. (1994). Understanding language disorders: The impact on learning. Eau Claire, WI: Thinking Publications.

Richards, R. G. (1999). The source for dyslexia and dysgraphia. East Moline, IL: LinguiSystems.

Rousseau, M. K. (1990). Errors in written language. In R. A. Gable & J. M. Hendrickson (Eds.) Assessing students with special needs (pp. 89-101). London: Longman Group.

Rutter, P. & Raban, B. (1982). The development of cohesion in children's writing: A preliminary investigation. First Language, 3 (7), 63-75.

Sax, G. (1997). Principles of educational and psychological measurement and evaluation (4th ed.). Belmont, CA: Wadsworth Publishing.

Schiffrin, D. (1994). Approaches to discourse. Cambridge: Blackwell.

School District #57 (1996). Guidebook for the use of curriculum based measurement in School District #57. Prince George, BC: School District #57.

Scott, C. (1988). Spoken and written syntax. In M. A. Nippold (Ed.) Later language development: Ages 9 through 19 (pp. 49-95). Boston: College-Hill Press.

Scott, C. (1991a). Learning to write: Context, form and process. In A. G. Kamhi & H. W. Catts (Eds.) Reading disabilities: A developmental language perspective (pp. 261-302). Boston: Allyn & Bacon.

Scott, C. (1991b). Problem writers: Nature, assessment and intervention. In A. G. Kamhi & H. W. Catts (Eds.) Reading disabilities: A developmental language perspective (pp. 303-344). Boston: Allyn & Bacon.

Silliman, E. R., Jimerson, T. L. & Wilkinson, L.C. (2000). A dynamic systems approach to writing assessment in students with language learning problems. Topics in Language Disorders, 20 (4), 45-64.

Silliman, E. R. & Wilkinson, L.C. (1994). Observation is more than looking. In G. P. Wallach & K. G. Butler (Eds.), Language learning disabilities in school-aged children and adolescents: Some principles and applications (pp. 145-173). Toronto: Maxwell Macmillan Canada.

Silliman, E. R., Wilkinson, L.C. & Hoffman, L. P. (1993). Documenting authentic progress in language and literacy learning: Collaborative assessment in the classrooms. Topics in Language Disorders, 14 (1), 58-71.

Singer, B. D. (1995). Written language development and disorders: Selected principles, patterns and intervention possibilities. Topics in Language Disorders, 16(1), 83-96.

Smith, L. (1999). An exploration of cohesion in narrative and expository writing in the mid-elementary years. Unpublished paper. (Available from Lynda Struthers [Smith], 267 Claxton Cres., Prince George, BC, V2M 5X6).

Stevens, J. (1996). Applied multivariate statistics for the social sciences (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Tabachnick, B. S. & Fidell, L. S. (1996). Using multivariate statistics. (3rd ed.). New York: Harper Collins College Press.

Thompson, S. (1997). The source for nonverbal learning disorders. East Moline, IL: LinguiSystems.

Tindal, G., Marston, D. & Deno, S. L. (1983). The reliability of direct and repeated measurement (Research Report No. 109). Minneapolis, MN: Institute for Research on Learning Disabilities.

Tindal, G. & Nolet, V. (1990). The construct validity of curriculum-based measurements of achievement: A multitrait-multimethod analysis. Paper presented at the American Educational Research Association, Boston, April 16-20. (ERIC Document Reproduction Service No. ED 325 506)

Tindal, G. & Parker, R. (1991). Identifying measures for evaluating written expression. Learning Disabilities Research and Practice, 6, 211-218.

Tinsley, H. E. & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgements. Journal of Counseling Psychology, 22 (4), 358-376.

Warren, S. F. & Yoder, P. J. (1994). Communication and language intervention: Why a constructivist approach is insufficient. The Journal of Special Education, 28 (3), 248-258.

Wiig, E. & Semel, E. (1984). Language assessment and intervention for the learning disabled (2nd ed.). Columbus, OH: Charles E. Merrill.

Zarnowski, M. (1981). A child's composition: How does it hold together? Language Arts, 58 (2), 316-19.

APPENDIX A

Letter of Consent

(Principal's Name)
(School's Name and Address)

March 10, 2000

Dear ,

I am currently working as a speech-language pathologist on Area Support Team 4. I am also currently working on completing graduate studies in Education at the University of Northern British Columbia (UNBC). This letter contains an outline of my thesis and requests your assistance in completion of this research. Norm Monroe, Director of School Services, has given his approval for this thesis project. He will be kept apprised of the details of this project as it is carried out. The results will be of interest to support teachers and school district specialists involved in student assessment.

Research Problem to be Addressed

Review of the literature and my own experience indicates that there are qualitative differences between the writing of children with disabilities and those without. These qualitative differences are not just a reflection of the misuse of writing conventions such as spelling, punctuation and grammar, but extend into the ability of writers to communicate their ideas effectively to the reader. The research indicates that problem writers have difficulty with cohesion (e.g. Hedberg & Fink, 1996). Cohesion involves the use of linguistic devices that serve to link ideas and sentences together creating a unified text. This area of writing has been shown to reflect the readability of a text (e.g. Crowhurst, 1980; Hedberg et al., 1996), but very few assessment tools are available that evaluate writing in this way. My research, therefore, involves the development of an assessment instrument that can be used to assess cohesion in the writing samples of school-aged children for the purpose of planning and monitoring interventions.

Method

The method I will use involves a field test of the instrument using approximately 300 to 400 CBM writing samples from students in School District #57, in Grades 4 through 7. I am requesting that some of these samples be from your school. The samples from your school should consist of all students in each grade in a single testing period. The time of year in which the samples were collected is not significant, though I ask that all samples provided come from the same testing period. The samples need not be current and names of the students and schools who generated them will remain anonymous. The only identifying information that is requested is the grade and gender of the writer, as well as any special designation that the student might have (SLR, ESL/ESD, LD). Photocopies or original samples are welcome. Originals will be returned at the completion of the project. Photocopies will be destroyed.

The collected samples will each be rated using the cohesion assessment tool. The results of this rating will then be used to test for sensitivity of individual test items, and the instrument's reliability and validity.

Ethical Considerations

The samples used in this research will have been generated for educational rather than research purposes. As the purpose of examining the samples is to evaluate my assessment tool rather than students, there is no perceived harm to individuals. Furthermore, as the samples will have all identifying information other than grade, gender, and special designation removed, the confidentiality of the school and student will be protected.

Summary

This thesis has grown out of an interest in developing assessment tools that are functional, and useful for developing and monitoring goals in educational intervention plans. The completion of this project will depend on your assistance in supplying CBM writing probes from your school as specified above. Please feel free to contact Norm Monroe or myself if you see any difficulty with the project as it is presented here.

Thank you in advance for your assistance and support.

Lynda Smith
Area Support Team # 4
562-3780
e-mail: Lynda_Smith@fc.schdist57.bc.ca

Norm Monroe
Director of School Services
561-6800 ext. 311
norm_monroe@fc.schdist57.bc.ca

References

Crowhurst, M. (1980). The effect of syntactic complexity on writing quality: A review of the research. Unpublished paper. ERIC document ED 202 024.

Hedberg, N. L. & Fink, R. J. (1996). Cohesive harmony in the written stories of elementary children. Reading and Writing: An Interdisciplinary Journal, 8, 73-86.

APPENDIX B

Procedures for Administering CBM Writing Probes

The directions for administration of the CBM writing probes as outlined in the Guidebook for the Use of Curriculum Based Measurement in School District # 57 (School District # 57, 1996) are listed here.

Materials

1. Story starter.
2. Stop watch.

Directions

1. Select an appropriate story starter.
2. Provide the student with a pencil and a sheet of lined paper.
3. Say these specific directions to the students:

   "You are going to write a story. First I will read a sentence, and then you will write a story about what happens next. You will have one minute to think about what you write, and three minutes to write your story. Remember to do your best work. If you don't know how to spell a word you should guess. Are there any questions? (Pause.) Put your pencils down and listen. For the next minute, think about... (insert story starter)."

4. After reading the story starter, begin your stopwatch and allow 1 minute for students to "think." (Monitor students so that they do not begin writing.) After 30 seconds say: "You should be thinking about... (insert story starter)."
5. At the end of 1 minute say: "Now begin writing." Restart your stop watch.
6. Monitor students' attention to the task. Encourage students to work only if they are looking around and talking.
7. After 90 seconds say: "You should be writing about (insert story starter)."

8. At the end of 3 minutes say: “Stop. Put your pencils down.

The three story starters used to generate the writing samples used in this study included:

1. Yesterday, a monkey climbed through the school and...

2. I was walking along a path when all of a sudden...

3. The cat climbed the telephone pole and...

APPENDIX C

Examples of Writing Samples

Included here are examples of the writing samples used in this study. These include examples of typical, best and worst writing samples produced by each grade. The typical examples shown here were chosen as their scores for fluency, syntactic complexity and cohesion reflected the mean for each score for a given grade. The worst and best examples were selected from those with the highest and lowest cohesion scores for each grade. Each sample is re-written here with the original spelling and punctuation used by the writer.

Grade 4

Typical. Yesterday a monkey climbed through the window at school and... the monkey came and pulled all the teachers hair off so she was bald and the monkey tore every ones paper and broke the chairs and desks and distrode the hole classroom and he did that...

Worst. Yesterday a monkey climbed through the window at school and... eat my banna I was so mad I got a 22 and shont him 100 they staid at me and...

Best. I was walking along a path when all of a sudden... a dog jumped on me and pushed me down and then it liked me. I was scared it would bite me. But it didn't. so I got up and took the dog home and showed it to my mom. She said...

Grade 5

Typical. Yesterday a monkey climbed through the window at school and... ate all our math books. After school our teacher, Mrs. White, took the monkey to the zoo and of course we got an F on Math. Today we were working on SS and a lion jumped through the window and ate our SS books.

Worst. The cat climbed the telephone pole and... stepped on to the thin wire. I have to get my baby she thought. Everyone around here were screaming, "here kitty kitty, come down for there, she'll jump,". She just...

Best. I was walking along a path when all of a sudden... a frog jumped out and I know how I hate frogs so I lept behind a bush until the frog left. When I left the bush I thought I was safe but just when I turned the corner there was over a million frogs. I had no way to get away and I think I squashed 11 frogs. When I finally made it over, I ran home and swore...

Grade 6

Typical. I was walking along a path when all of a sudden... I saw a poor hurt baby panther that had a thorn threw its paw. I went close but not too close. Then its mother came and started to growl at me. I backed away and the baby limped closer towards me. I was frozen stiff.

Worst. Yesterday a monkey climbed through the window at school and... grabbed a experiment and smashed it. It blew the room up. The monkey grabbed more things and smashed them. There was these toxic chemicals that burnt the school down. All the kids were screaming. The monkey wasn't smart enough to move and died. The fireman came a put the fire out. They told the kids they...

Best. Yesterday a monkey climbed through the window at school and... it started throwing everything it could at people. First it through the encyclopedias then the dictionaries then the text books. When the zoo keepers came it through chairs and pencils, anything it could find at the zookeepers but finally they caught it but the zookeepers all had bumps and bruises. All this took...

Grade 7

Typical. Yesterday a monkey climbed through the window at school and... started jumping around crazily. The kids thought that if was really cool but the teacher jumped up, started screaming and ran down the hall. The monkey started swinging from the roof and soon dropped onto my desk, at first it frightened me but then I realized that he did not want to hurt me so I calmed down. Soon the principal...

Worst. I was walking along a path when all of a sudden... the school bully came. "Oh no" I pannicked. "Help" He stopped me in my tracks. "Lunch mony.", he demanded I quickly thought up a lie, no he knows I have money. "uh,um No" I whispered "No!" He boomed "No, did you say No."

Best. I was walking along a path when all of a sudden... I was in a jungle wear lions, tigers, and bears and many other wild animals live when a lion chased me. I started to freak out and screamed but I thought that wouldn't do much so I ran and ran until I saw the lion was not behind me anymore and when I stopped I saw some strawberries so I ate them and when they were gone I got tired. So I fell asleep. When I woke up I was in tarzans little treehouse up really high. So then I saw...

APPENDIX D

Checklist 1.0

	YES	NO
Cohesive marker	1	0
1. All pronouns refer to some previously mentioned noun.		
2. All pronouns have a referent in the previous sentence.		
3. All demonstratives (eg. this, these, that, those) have a clear referent in the previous sentence.		
4. All nouns appearing with the article 'the' have a previous referent in the text.		
5. Referents for nouns used with 'the' that are not present in the text can be inferred from world knowledge.		
6. Each sentence is connected to the one preceding it by at least one anaphoric reference.		
7. The written passage is sequentially organized.		
8. ‘And' is used to connect sentences and/or clauses.		
9. ‘Also' is used to connect sentences and/or clauses.		
10. Other coordinating conjunctions are used to connect sentence and/or clauses.		
11. 'Then' is used to connect sentences and/or clauses.		
12. ‘When’ is used to connect clauses.		
13. ‘Before' and 'after' are used to connect sentences and/or clauses.		
14. ‘First, second,' etc. are used to connect sentences and/or clauses.		
15. Other temporal conjunctions are used to connect sentences and/or clauses.		
16. Consistent tenses are used throughout.		
17. Shifts in time are marked with temporal terms (eg. the next day) other than conjunctions.		
18. Causal relationships are implied.		
19. ‘So' is used to connect sentences and/or clauses.		
20. ‘Because' is used to connect clauses.		
21. Other causal conjunctions are used to connect sentences and/or clauses (eg. consequently, therefore, etc.).		
22. Adversative conjunctions (eg. but) are used to connect clauses.		
23. Super-ordinates, synonyms or near-synonyms are used for the same referent in adjacent sentences.		
24. Complementary terms, converses or antonyms appear in adjacent sentences.		
25. The text is divided into paragraphs.		
26. Paragraphs have topic sentences.		
27. Explicit transitions between paragraphs are present.		
Total Cohesion Score		

APPENDIX E

Checklist 1.1

	YES	NO
Cohesive marker	1	0
1. All pronouns refer to some previously mentioned noun.		
2. All pronouns have a referent in the previous sentence.		
3. All demonstratives (eg. this, these, that, those) have a clear referent in the previous sentence.		
4. Referents for nouns used with 'the' that are not present in the text can be inferred from world knowledge.		
5. All nouns appearing with the article 'the' have a previous referent in the text.		
6. Each sentence is connected to the one preceding it by at least one anaphoric reference.		
7. The written passage is sequentially organized.		
8. 'And' is used to connect sentences and/or clauses.		
9. 'Also' is used to connect sentences and/or clauses.		
10. Other coordinating conjunctions are used to connect sentence and/or clauses.		
11. 'Then' is used to connect sentences and/or clauses.		
12. 'When' is used to connect clauses.		
13. 'Before' and 'after' are used to connect sentences and/or clauses.		
14. 'First, second,' etc. are used to connect sentences and/or clauses.		
15. Other temporal conjunctions are used to connect sentences and/or clauses.		
16. Consistent tenses are used throughout.		
17. Shifts in time are marked with temporal terms (eg. the next day) other than conjunctions.		
18. Causal relationships are implied.		
19. 'So' is used to connect sentences and/or clauses.		
20. 'Because' is used to connect clauses.		
21. Other causal conjunctions are used to connect sentences and/or clauses (eg. consequently, therefore, etc.).		
22. Adversative conjunctions (eg. but) are used to connect clauses.		
23. Super-ordinates, synonyms or near-synonyms are used for the same referent in adjacent sentences.		
24. Complementary terms, converses or antonyms appear in adjacent sentences.		
25. The text is divided into paragraphs.		
26. Paragraphs have topic sentences.		
27. Explicit transitions between paragraphs are present.		
Total Cohesion Score		

APPENDIX F

Checklist 1.2

	YES	NO
Cohesive marker	1	0
1. All pronouns refer to some previously mentioned noun.		
2. All pronouns have a referent in the previous sentence.		
3. All demonstratives (eg. this, these, that, those) have a clear referent in the previous sentence.		
4. Referents for nouns used with 'the' that are not present in the text can be inferred from world knowledge.		
5. All other referents for nouns used with the article 'the' have a previous referent in the text.		
6. Each sentence is connected to the one preceding it by at least one anaphoric reference.		
7. The written passage is sequentially organized.		
8. 'And' is used to connect sentences and/or clauses.		
9. 'Also' is used to connect sentences and/or clauses.		
10. Other coordinating conjunctions are used to connect sentence and/or clauses (eg. or, another, as well as, etc.).		
11. 'Then' is used to connect sentences and/or clauses.		
12. 'When' is used to connect clauses.		
13. 'Before' and 'after' are used to connect sentences and/or clauses.		
14. 'First, next,' etc. are used to connect sentences and/or clauses.		
15. Other temporal conjunctions are used to connect sentences and/or clauses.		
16. Consistent tenses are used throughout.		
17. Shifts in time are marked with temporal terms (eg. the next day) other than conjunctions.		
18. Causal relationships are implied.		
19. 'So' is used to connect sentences and/or clauses.		
20. 'Because' is used to connect clauses.		
21. Other causal conjunctions are used to connect sentences and/or clauses (eg. consequently, therefore, etc.).		
22. Adversative conjunctions (eg. but) are used to connect clauses.		
23. Super-ordinates, synonyms or near-synonyms are used for the same referent in adjacent sentences.		
24. Super-ordinates, synonyms or near-synonyms are used for the same referent across the text.		
25. Complementary terms appear in adjacent sentences.		
26. Converses or antonyms appear in adjacent sentences.		
27. A new paragraph is used when there is a shift in story events.		
Total Cohesion Score		

APPENDIX G

Checklist 2.0 With Instruction Manual

COHESION CHECKLIST

Developed by

Lynda Smith

April 14, 2000

Revisions: August, 2000

Table of Contents

Background

This checklist was designed to evaluate the linguistic elements used to achieve cohesion in the writing of elementary school-aged children. Cohesion consists of the ties that link sentences and ideas together to form a unified, single text (Halliday & Hasan, 1976), that is comprehensible to the reader (Hedberg & Fink, 1996; Lindberg, 1984; Zarnowski, 1981). Without it, writing would consist of a series of unrelated sentences or ideas.

The content of the checklist originates from several studies of cohesion (Crowhurst, 1981, 1987; Halliday et al., 1976; Liles, 1985; McCutchen and Perfetti, 1982; Scott, 1991; Smith, 1999). The aspects of cohesion examined through the items on this checklist include reference, conjunction, lexical cohesion and overall global structures.

Reference includes the use of pronouns, articles and demonstratives to refer to information within the text. Conjunction is used to connect clauses and sentences and to organize text. The conjunctions evaluated here include additive (eg. and), temporal (eg. then), causal (eg. because), and adversative (eg. but) forms. Lexical cohesion is accomplished through reiteration of a term using the same word, a superordinate, a synonym or near-synonym, or collocation which involves use of words that commonly occur together such as antonyms, complementary terms and converses. The degree of cohesion accomplished through lexical reiteration and collocation is a reflection of how close the words are in meaning and the distance between them within the written text. The degree of cohesion is stronger where the distance is less. Global structures that affect cohesion include consistencies used across a text, such as tense marking, and overall organization of a piece, such as temporal organization, causal relationships and paragraph structure.

Definition of Key Terms

adversative - marking an opposing or contrary relationship antonyms - words that mean the opposite of each other. An example would be 'hot' and 'cold'. These word pairs affect cohesion because of their strong semantic relationship.

clause (clausal) - a clause consists of a subject and verb. Clauses may be independent, in which case they can stand alone as a sentence. They may also be subordinating, in which case they need to be attached to an independent clause by a subordinating conjunction to complete the thought. Subordinating clauses consist of those which begin with conjunctions such as 'because', 'when', 'until', or 'although'. Independent clauses may be joined to other independent clauses using coordinating conjunctions such as 'and', 'or', or 'but'. 'So' and 'then' are often also treated as coordinating conjunctions in narrative analysis (Hughes, McGuillivray & Schmidek, 1997).

complementary terms - words that often appear together and thus complement one another. Such word pairs consist of terms that have associated meanings such as 'joke' and 'laugh,' or 'lake' and 'beach'. Such terms are important to cohesion due to their strong semantic relationship.

converses - these consist of word pairs that reflect a relationship of response of one term to the other. Examples of converses include 'lead' and 'follow' or 'throw' and 'catch'. Closely related to antonyms, converses are important to cohesion due to their strong semantic relationship.

lexical - relating to words or the semantic relationship between words. This reflects word meaning.

near-synonym - words that are used to refer to the same thing, but may not have identical meanings when used out of context. Examples of near-synonyms include 'lion' and 'beast,' or 'cave' and 'shelter'. Such uses of near-synonyms are important to cohesion due to their strong semantic relationship and common reference.

reiteration - mentioning a person/place/thing/idea more than once in the same written text through direct repetition of a word, or replacement with a word that refers to the same thing.

sentence - a sentence consists of an independent clause containing a subject and a verb and any attached subordinate clauses. For the purposes of this evaluation, a sentence need not be signaled by mechanical conventions such as capital letters and periods. The boundaries of the sentence are determined by the subject/verb parameters mentioned above.

super-ordinate - a categorical label that can be used to replace a more specific term. For example, 'animal' is the super-ordinate of 'dog'.

synonym - a different word used to mean the same thing. An example is 'car' and 'automobile'.

Scoring Instructions

The following descriptions provide the criteria for scoring the corresponding items on the checklist. The sum of the scores for all 25 items will provide the total cohesion score. Composite scores will be derived from each section.

Before scoring the writing sample for cohesion, first mark the boundaries between *sentences* (see above definition) to clarify the beginning and ending of independent clauses. For more information on dividing samples in this manner see Hughes et al. (1997). Read through the entire sample once to familiarize yourself with the content before going through the items on the checklist. Ignore missing words as though they were purposely omitted. For example, if a child missed an article before a noun, do not treat it as a possible credit for 'the' on the checklist. Similarly, do not treat missing words as examples of errors. In this manner, if a child misses an article but uses 'the' correctly in every other instance, he or she would receive credit for the item. Treat incomplete thoughts as independent clauses or sentences.

When scoring items for reference, a sore of one is achieved by demonstrating use as described on the checklist in **all** cases. For conjunction and lexical cohesion, only one example need be present in the sample to receive credit. Also, conjunctions must be used to join sentences or clauses and will not be given credit when used elliptically (eg. Someday I'll go to the moon. I don't know when.) Two conjunctions used together (eg. and then...) receive two credits.

Item # Scoring Criteria

Reference

1. Score 1 if every pronoun used refers unambiguously to a noun previously mentioned in the text. Score 0 if any pronoun has more than one possible referent, or no referent mentioned in the text. The first appearance of first and second pronouns 'I' and 'you' are treated as nouns. Subsequent uses will be treated as pronouns. Disregard uses of 'it' that are used to establish setting (eg. It was a warm sunny day.). Also score 0 if no pronouns are used in the sample.

2. Score 1 if every pronoun refers unambiguously to a noun or pronoun **in the previous sentence or clause**. Score 0 if any referent is not unambiguously contained in the previous or same sentence. Apply the rules for uses of 'I', 'you' and 'it' established in item #1.

3. Score 1 if every demonstrative (this, that, these and those) is used with an object/person/place/idea that has been mentioned in the previous sentence. This previous mention of a noun need not be an exact repetition of the same word but must refer to the same thing. Score 0 if the object/person/place/idea was mentioned somewhere other than the previous sentence or not at all. Also score 0 if no demonstratives are used.

4. Score 1 if every occurrence of 'the' is used with a noun that has an unambiguous referent. It should be clear to the reader to which specific person, place, thing or idea the writer is referring.. Occurrences of the word 'the' may be used next to a noun that has a previous mention in the text. This mention may include reference to the same object/person/place/idea using a different word. For example, the following use of 'the' would qualify for a score of 1:

I saw a dog running toward me. The beast looked mean.

'The' may also be used to refer to a special case so that the referent can be inferred from the content of the text or world knowledge. 'The' may also be used to introduce setting. The following examples would also qualify for a score of 1:

I saw the Prime Minister on TV.

We live on the earth.

I walked into a store and asked to speak to the manager.

The day was warm and sunny.

Score 0 if the referent for a noun used with 'the' cannot be unambiguously inferred from context, world knowledge, or previous mention in the text. Also score 0 if 'the' is not used.

5. Score 1 if **every sentence** contains a reference through the use of pronouns, demonstratives or the definite article 'the' **to the sentence directly preceding it**. In this case, to qualify for credit, uses of 'the' must refer to something specifically mentioned in the previous sentence. Score 0 if any sentences in the text do not refer directly to elements of the sentence preceding it.

Conjunction

6. Score 1 if the conjunction 'and' is used to join any two independent clauses. Score 0 if 'and' is not present in the written text or if it is only present to create compound subjects, or verb phrases. For example, the following uses of 'and' would not receive credit on this item:

The boy and girl were running.

The ball was red and black.

The children were laughing and playing.

7. Score 1 if the conjunction 'also' is used to join any two independent clauses. Score 0 if 'also' is not present in the written text or if it is only present to create compound subjects, or verbs. For example, the following use of 'also' would not receive credit:

The boy and also the girl were hungry.

8. Score 1 if there is any indication of additive conjunctions being used to join any two independent clauses. Examples of additive conjunctions include "another, or, in addition/additionally, as well as, etc." A semi-colon may also be used in this fashion and if used correctly would score 1. Score 0 if there are no other additive conjunctions used besides 'and' and 'also'.

9. Score 1 if the term 'then' is used to join or relate any two independent clauses. Score 0 if 'then' is not present in the written text.

10. Score 1 if the term 'when' is used to join or relate any two clauses. Score 0 if 'when' is not present in the written text.

11. Score 1 if the term(s) 'before' or 'after' are used to join or relate any two clauses. It is not necessary for both of these terms to be present to receive credit. Either one will warrant a score of 1. Score 0 if these terms are not present in the writing sample.

12. Score 1 if any subordinating temporal conjunctions other than the ones mentioned above are used to join or relate any two clauses. These may include terms like "until, while, as, ...". Score 0 if no other subordinating temporal conjunctions are present in the written passage. Caution: use of the word 'as' must reflect a temporal rather than causal meaning to receive credit (eg. I gazed at the horizon as the moon was setting.).

13. Score 1 if **adverbs or adverbial phrases** are used to mark shifts in time or the sequence of events. These might include temporal terms like "first, next, finally...", their adverbial derivatives (eg. firstly), or phrases such as "all of a sudden, the next

day, later on, the following week, etc.". Sequential markers like 'first, second, last, etc.' need not be presented in a series to receive credit. The following examples would receive a score of 1:

<u>First</u> of all, the boy ate his hotdog. He ate some candy <u>next</u>.

OR

<u>First</u> the boy was frightened. Then he got mad.

Score 0 if no such adverbs or phrases appear in the text.

14. Score 1 if the conjunction 'so' is used to join any two independent clauses. Score 0 if 'so' is not present in the written text.

15. Score 1 if the conjunction 'because' is used to join any two clauses. Score 0 if 'because' is not present in the written text.

16. Score 1 if any other causal conjunctions are used to join clauses. Examples of additive conjunctions include "consequently, therefore, etc." . Score 0 if there are no other causal conjunctions used other than "so" and "because".

17. Score 1 if conjunctions showing an adversative relationship (eg. but, however, although) are used to join or relate any two clauses . Score 0 if no such conjunctions appear.

Lexical Cohesion

18. Score 1 if any referent is reiterated in an adjacent sentence through the use of superordinate, synonym, or near-synonym terms. An example of superordinates might be word pairs like 'dog' and 'animal'. Synonyms involve the use of word pairs that mean the same thing like 'dog' and 'canine'. Near-synonyms consist of word pairs with similar meanings that refer to the same thing. For example the following pairs of sentences contain an example of a near-synonym:

"He held a knife in his hand. He waved the blade wildly."

The referent of the term must be clear to the reader. Score 0 if reiteration consists only of repetition of the same word from sentence to sentence or if it does not occur at all.

19. Score 1 if any referent is reiterated through use of super-ordinate, synonym, or near-synonym terms across the text. In this case, credit is given for such terms **not** occurring in adjacent sentences. Score 0 if reiteration across the text consists only of repetition of the same word or if there are no examples of super-ordinates, synonyms, or near-synonyms across the text.

20. Score 1 if any word pairs with complementary semantic links appear in neighboring sentences. Complementary terms include words that commonly occur together like 'boy-girl' or 'play-fun'. Such terms may also reflect topic maintenance by referring to things that commonly occur together. The following sentence pairs reflect this type of semantic connection:

He fired the gun. A bullet grazed my ear.

The UFO landed. Aliens appeared.

Score 0 if complementary word pairs do not appear in neighboring sentences anywhere in the passage.

21. Score 1 if any word pairs with semantic links such as converses or antonyms appear in neighboring sentences. Converses include items that suggest a response of one to the other. These might include terms such as 'order-obey' or 'listen-tell'. An example would be:

He spoke. I listened.

Score 0 if no such lexical pairs appear in neighboring sentences anywhere in the passage.

Global Organization

22. Score 1 if the written text has a general sequential order (ie. things that happened first are mentioned first, etc.). Score 0 if the text consists of randomly ordered ideas or does not flow in a temporal sequence.

23. Score 1 if one tense (eg. past, present, or perfect) is used consistently throughout the passage. A score of 1 would also apply if shifts in tense occur in passages of dialogue. If a passage were written in past tense with a quotation written in another tense, it would still receive a score of 1. For example:

> The girl ran down the hall. She shouted to the people standing there, " Will you help me?"

Score 0 if tenses are used inconsistently in the writing sample.

24. Score 1 if any event in the passage is causally linked to another event mentioned in the previous sentence. The two events need not be explicitly linked to receive credit. For example, the sentences "I was hungry. I went inside to get something to eat." show an implicit causal connection. The sentences "I was running. I stopped." does not imply or demonstrate a causal connection. A score of 0 applies when no causal links between adjacent sentences are detected.

25. Score 1 if a new paragraph is used when there is a shift in the story's events. These might include introduction of a new speaker, a new location or a new time. The paragraph need not be indented but should be marked by a new line of writing. Score 0 if there is only one paragraph in the sample or if new paragraphs are not introduced with changes in speakers, location or time.

Table 1A indicates the break down of cohesive devices examined by this instrument. Each item from the checklist is listed in the table next to the cohesion category(s) it represents.

Table 1A

Table of Test Specifications

Type of Cohesive Device		Corresponding Item Numbers
Referential Cohesion		1, 2, 3, 4, 5
Conjunction:	additive	6, 7, 8
	temporal	9, 10, 11, 12, 13
	causal	14, 15, 16
	adversative	17
Lexical Cohesion		3, 4, 18, 19, 20, 21
Global Organization		22, 23, 24, 25

References

Crowhurst, M. (1981) Cohesion in argumentative prose written by sixth-, tenth-, and twelfth-graders. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles.

Crowhurst, M. (1987). Cohesion in argument and narration at three grade levels. Research in the Teaching of English, 21(2), 185-197.

Halliday, M.A.K. & Hasan, R. (1976). Cohesion in English. London: Longman Group.

Hedberg, N.L. & Fink, R.J. (1996). Cohesive harmony in the written stories of elementary children. Reading and Writing: An Interdisciplinary Journal, 8, 73-86.

Hughes, D., McGillivray, L & Schmidek, M. (1997). Guide to narrative language: Procedures for assessment. Eau Claire, WI: Thinking Publications.

Liles, B.Z. (1985). Cohesion in the narratives of normal and language-disordered children. Journal of Speech and Hearing Research, 28, 123-133.

Lindeberg, A.C. (1984). Cohesion and coherence in short expository essays. Proceedings from the Nordic Conference for English Studies.

McCutchen, D. & Perfetti, C.A. (1982). Coherence and connectedness in the development of discourse production. Text, 2, 113-139.

Scott, C. (1991a). Learning to write: Context form and process. In A.G. Kamhi & H.W. Catts (Eds.) Reading disabilities: A developmental language perspective (pp. 261-302). Boston: Allyn & Bacon.

Smith, L. (1999). An exploration of cohesion in narrative and expository writing in the mid-elementary years. Unpublished paper. (Available from Lynda Struthers, 267 Claxton Cres., Prince George, BC, V2M 5X6).

Zarnowski, M. (1981). A child's composition: How does it hold together? Language Arts, 58 (2), 316-19.

Cohesive marker	YES	NO
1. All pronouns refer to some previously mentioned noun.	1	0
2. All pronouns have a referent in the previous sentence or clause.	1	0
3. All demonstratives (eg. this, these, that, those, here, there) have a clear referent in the previous text.	1	0
4. Referents for nouns used with 'the' have an unambiguous previous mention in the text or can be inferred from world knowledge.	1	0
5. Each sentence is connected to the one preceding it by at least one form of reference.	1	0

REFERENCE SUB-SCORE

6. ‘And' is used to connect independent clauses.	1	0
7. ‘Also' is used to connect independent clauses.	1	0
8. Other coordinating conjunctions are used to connect independent clauses (eg. or, another, as well as, etc.).	1	0
9. ‘Then' is used to connect independent clauses.	1	0
10. ‘When’ is used to connect clauses.	1	0
11. 'Before' or 'after' is used to connect clauses.	1	0
12. Other subordinating temporal conjunctions are used to connect clauses.	1	0
13. Adverb or adverbial phrases are used to mark sequence or shifts in time (eg. First, last, all of a sudden, etc.)	1	0
14. ‘So’ is used to connect sentences and/or clauses.	1	0
15. ‘Because' is used to connect clauses.	1	0
16. Other causal conjunctions are used to connect sentences and/or clauses (eg. consequently, therefore, etc.).	1	0
17. Adversative conjunctions (eg. but) are used to connect clauses.	1	0

CONJUNCTION SUB-SCORE

18. Super-ordinates, synonyms or near-synonyms are used for the same referent in adjacent sentences.	1	0
19. Super-ordinates, synonyms or near-synonyms are used for the same referent across the text.	1	0
20. Complementary terms appear in adjacent sentences.	1	0
21. Converses or antonyms appear in adjacent sentences.	1	0

LEXICAL COHESION SUB-SCORE

22. The written passage is sequentially organized.	1	0
23. Consistent tense is used throughout.	1	0
24. A causal relationship exists between two adjacent sentences .	1	0
25. A new paragraph is used when there is a shift in story events.	1	0

GLOBAL ORGANIZATION SUB-SCORE

Total Cohesion Score ______

Scoring Summary Sheet

Student's Name: ______________________________

Examiner's Name: ____________________________

Date of Writing Sample: ________________________

Grade: ______________________________________

Referential Cohesion ___________

Conjunction ___________

Lexical Cohesion ___________

Global Organization ___________

Total Cohesion Score ___________

Scoring Companion

Reference

Item	**Scoring Criteria**	**Example**	**Non-example**
1	**Every** pronoun refers to a previously mentioned noun, 'I' or 'you'.	the referent for every pronoun is clear	a pronoun with no referent or the referent is unclear
2	Every pronoun refers to a noun or pronoun **in the previous or same sentence.**	*The boy was walking fast. He was headed home.* OR *The man was afraid so he ran.* OR *She walked. She laughed.* OR *She cried. Tears rolled down her cheeks.*	referent not in the previous sentence or clause *The girl ate some chips. The chips tasted very good. Then she drank her pop.* OR *The girl waved her hand.*
3	Every noun used with a demonstrative has mention in the previous sentence.	uses *this, that, those* or *these* before a noun *A dog came running at us. That beast was mean.*	*The dogs came running at us. We were scared. Those dogs were mean.* OR Use of a demonstrative with a noun with no previous mention.
4	'The' is used with a noun with an unambiguous referent previously mentioned or understood from world knowledge.	*the Prime Minister* *the manager* *The day was warm and sunny.* *A dog chased me up the street. I ran as fast as I could. The dog looked mean.*	'the' used with the first mention of a noun that is not a special case: *the cat* *the officer* or 'the' used when it is not clear to the reader which specific person/place/thing/idea is being referred to
5	**Every sentence** contains a reference to the **previous sentence**	reference to previous sentence through the correct use of *pronouns, demonstratives*, or *the definite article 'the'* *(use of 'the' must indicate something mentioned in the previous sentence)*	At least one sentence is not connected to the previous one by use of *pronouns, demonstratives*, or *the definite article 'the'*

Conjunction

Item	**Scoring Criteria**	**Example**	**Non-example**
6	Use of 'and' to connect independent clauses	*The children were playing and the boys were running.* OR *The dog was mean. And I was afraid.*	*The boy and girl were running.* *The ball was red and black.* *The children are laughing and playing.* OR 'and' not used
7	Use of 'also' to connect independent clauses	*I was late. I also was hungry.* OR *I ate some cheese and I also ate some crackers.*	*The boy and also the girl were hungry.* OR 'also' not used
8	Use of additive conjunctions other than 'and' and 'also'	*or, another, in addition, additionally, as well as, furthermore, besides, nor* etc.	no other additive conjunctions
9	Use of 'then'	*I went to the store. Then I went home.*	no use of 'then'
10	Use of 'when'	*When I got home I ate lunch.*	no use of 'when'
11	Use of 'before' or 'after' to connect clauses	*I did my homework before I went outside.* OR *After I ate lunch, I went home.*	no use of 'before' or 'after' OR 'before' and 'after' not connecting clauses *I was there before.*
12	Use of other subordinating temporal conjunctions	*until, while, as, since(time),* etc.	no other temporal conjunctions than those listed in items 11-14
13	**Adverbs or adverbial phrases** used to mark shifts in time or sequence	*first, second, third, next, last, finally, suddenly, later, all of a sudden, as soon as, the next day, later on, the following week, after that,* etc	no adverbs or phrases used marking shifts in time
14	Use of 'so'	*I was hungry. So I ate something.* OR *I was late so I hurried home.*	'so' used in a way that conveys degree rather than causation *I was so hungry, I could eat a horse.* OR 'so' is not used

Item	Scoring Criteria	Example	Non-example
15	Use of 'because'	*Because I was sad, I went home.* OR *I laughed because I was so happy.* OR *I ran. Because I was in a hurry.*	'because' is not used OR 'because' does not connect two ideas *I went home because.*
16	Use of other causal conjunctions	*consequently, therefore, since (cause),* etc.	no other causal conjunctions used other than 'so' and 'because' OR none used at all
17	Use of adversative conjunctions	*but, however, although, yet, instead, except, though,* etc.	no adversative conjunctions used

Lexical Cohesion

18	Uses reiteration of a referent in **adjacent sentences** at least one time (applies to nouns only)	superordinates *I saw a dog. The animal was huge.* OR synonyms *I saw a dog. The mutt was huge.* OR near-synonyms *I saw a dog. The beast was huge.* *note- these items will be signaled by the use of the definite article 'the' or a demonstrative.	repetition of a word *I saw a dog. The dog was huge.* OR superordinates, synonyms, or near-synonyms are used but **not in adjacent sentences**. OR No superordinates, synonyms, or near-synonyms are used
19	Uses reiteration of a referent at least one time **across the text** (applies to nouns only)	use of a synonym, near-synonym, or super-ordinate as described above but not in adjacent sentences	repetition of the same word OR reiteration only in adjacent sentences OR no use of synonyms, near-synonyms, or super-ordinates

Item	Scoring Criteria	Example	Non-example
20	Use of complementary terms in **adjacent sentences**	*The gun fired. A shot rang out.* OR *We went to the beach. The sand was hot.*	such terms are not used in adjacent sentences
21	Use of converses or antonyms in **adjacent sentences**	*speak-listen, ask-answer, order-obey, throw-catch, act-react,* etc. *You tell the story. I will listen.* OR *She climbed up. I slid down.*	such terms are not used in adjacent sentences

Global Organization

22	Events are mentioned in the order in which they occur.	sequence of events makes sense	sequence of events does not make sense
23	Consistent tense use	use of a single tense throughout the passage OR tense changes only in dialogue	inconsistent use of tense throughout the passage or tense mixing
24	Causal link between any two events in adjacent sentences	*I was hungry. I went inside to eat.* OR *I was crying because I lost my ball.*	one event may be related to but does not cause the other. *I was running. I stopped.*
25	New paragraphs used with new speakers, new time and new location.	a new line is started with each new speaker OR where there are shifts in time and place *The next day...* *Inside the house...*	no paragraph or no new line with a new speaker, time or location.

APPENDIX H

Checklist 2.1 With Modified Scoring Criteria

Cohesion Checklist

Student's Name: ______________________________ Grade: ____________________

Examiner's Name: ____________________________ School: ____________________

Date of Writing Sample: ________________________

Cohesive marker	YES	NO
1. All 3rd person pronouns refer unambiguously to some previously mentioned noun.	1	0
2. All 3rd person pronouns have an unambiguous referent in the previous sentence or same sentence.	1	0
3. All demonstratives (eg. this, these, that, those, here, there) have a clear referent in the previous text.	1	0
4. Referents for nouns used with 'the' have an unambiguous previous mention in the text or can be inferred from world knowledge.	1	0
5. Each sentence is connected to the one preceding it by at least one form of reference.	1	0

REFERENCE SUB-SCORE ______

6. Additive conjunctions are used to join independent clauses (e.g. and, also, or, another, as well as, etc.).	1	0
7. 'When' is used to connect clauses.	1	0
8. Other subordinating temporal conjunctions are used to connect clauses (e.g. before, after, until, while, as, etc.).	1	0
9. Adverb or adverbial phrases are used to mark sequence or shifts in time (e.g. then, next, first, last, all of a sudden, etc.).	1	0
10. Causal conjunctions are used to connect sentences and/or clauses (e.g. so, because, consequently, therefore, etc.).	1	0
11. Adversative conjunctions (eg. but) are used to connect clauses.	1	0

CONJUNCTION SUB-SCORE ______

12. Super-ordinates, synonyms or near-synonyms are used for the same referent in adjacent sentences or across the text.	1	0
13. Complementary terms appear in adjacent sentences.	1	0

LEXICAL COHESION SUB-SCORE ______

Total Cohesion Score ______

Item # Scoring Criteria

Reference

1. Score 1 if every 3rd person pronoun used refers unambiguously to a noun previously mentioned in the text. Score 0 if any pronoun has more than one possible referent, if different pronouns are used to refer to the same referent, or if an incorrect pronoun is used. Also score 0 if no referent for the pronoun is mentioned in the text. The following examples would result in a score of 0:

   The dog went home. It drank from its bowl. Then he ate.
   The girl dropped the ball. He tried to pick it up.
   John went to see Bill. He was in the school.

   Disregard uses of first and second person pronouns 'I' and 'you'. Disregard uses of 'it' that are used to establish setting (eg. It was a warm sunny day.). Also score 0 if no third person pronouns are used in the sample.

2. Score 1 if every third person pronoun refers unambiguously to a noun or pronoun **in the same or previous sentence**. **At least one** pronoun reference must **cross a sentence or clause boundary** to receive credit on this item. Score 0 if any referent is not unambiguously contained in the previous or same sentence. Also score 0 if all referents are contained in the same clause as their pronouns. Apply the rules for uses of 'I', 'you' and 'it' established in item #1.

3. Score 1 if every demonstrative (this, that, these, those, here and there) is used with or replaces an object/person/place/idea that has been previously mentioned in the text. This previous mention of a noun need not be an exact repetition of the same word but must refer to the same thing. Score 0 if the object/person/ place/idea was not mentioned in the previous sentence or if the referent is ambiguous. Disregard uses of

'there' to establish setting (eg. There were 12 people in the garden.). Also score 0 if no demonstratives are used.

4. Score 1 if every occurrence of 'the' is used with a noun that has an unambiguous referent. It should be clear to the reader to which specific person, place, thing or idea the writer is referring. **At least one** occurrence of the word 'the' should be used next to a noun that has a **previous mention in the text**. This mention may include reference to the same object/person/place/idea using a different word. For example, the following use of 'the' would qualify for a score of 1:

   I saw a dog running toward me. The beast looked mean.

   'The' may also be used to refer to a special case so that the referent can be inferred from the content of the text or world knowledge. 'The' may also be used to introduce setting. The following examples would also qualify for a score of 1:

   I saw the Prime Minister on TV.

   We live on the earth.

   I walked into a store and asked to speak to the manager.

   The day was warm and sunny.

   Score 0 if the referent for a noun used with 'the' cannot be unambiguously inferred from context, world knowledge, or previous mention in the text. Also score 0 if 'the' is never used to refer to a referent previously mentioned in the text or if 'the' is never used.

5. Score 1 if **every sentence** contains a reference through the use of pronouns, demonstratives or the definite article 'the' **to the sentence directly preceding it**. In this case, to qualify for credit, uses of 'the' must refer to something specifically

mentioned in the previous sentence. Score 0 if any sentences in the text do not contain a direct reference to elements of the sentence preceding it.

Conjunction

6. Score 1 if an additive conjunction is used to join any two independent clauses. Examples of additive conjunctions include "and, also, another, or, in addition/additionally, as well as, etc." A semi-colon may also be used in this fashion and, if used correctly, would score 1. Score 0 if additive conjunctions are not present in the written text or if they are only present to create compound subjects or verb phrases. For example, the following uses of 'and' would <u>not</u> receive credit on this item:

    The boy <u>and</u> girl were running.

    The ball was red <u>and</u> black.

    The children were laughing <u>and</u> playing.

7. Score 1 if the term 'when' is used to join or relate any two clauses. Score 0 if 'when' is not present in the written text or is used in a way that does not connect clauses.

8. Score 1 if subordinating temporal conjunctions other than 'when' are used to join or relate any two clauses. These may include terms like "before, after, until, while, as,...". Score 0 if these terms are not present in the writing sample or are not used to connect clauses. Caution: use of the word 'as' must reflect a temporal rather than causal meaning to receive credit (eg. I gazed at the horizon as the moon was setting.).

9. Score 1 if **adverbs or adverbial phrases** are used to mark shifts in time or the sequence of events. These might include temporal terms like "then, next, first, last, ...", their adverbial derivatives (eg. firstly,), or phrases such as "all of a sudden, the next day, later on, the following week, etc.". Sequential markers like "first, second, last,

etc." need not be presented in a series to receive credit. The following examples would receive a score of 1:

First of all, the boy ate his hotdog. He ate some candy next.

OR

First the boy was frightened. He got mad and went home.

Score 0 if no such adverbs or phrases appear in the text.

10. Score 1 if any causal conjunctions are used to join sentences or clauses. Examples of causal conjunctions include "so, because, consequently, therefore, etc." . Score 0 if there are no other causal conjunctions used or if they are used in a way that does not connect sentences or clauses.

11. Score 1 if conjunctions showing an adversative relationship (eg. but, however, although) are used to join or relate any two clauses . Score 0 if no such conjunctions appear or if they are used in a way that does not connect clauses.

Lexical Cohesion

12. Score 1 if any referent is reiterated anywhere in the text through the use of super-ordinate, synonym, or near-synonym terms. An example of superordinates might be word pairs like 'dog' and 'animal'. Synonyms involve the use of word pairs that mean the same thing like 'dog' and 'canine'. Near-synonyms consist of word pairs with similar meanings that refer to the same thing. For example the following pairs of sentences contain an example of a near-synonym:

"He held a knife in his hand. He waved the blade wildly."

The referent of the term must be clear to the reader. Score 0 if reiteration consists only of repetition of the same word from sentence to sentence or if it does not occur at all.

13. Score 1 if any word pairs with complementary semantic links appear in neighboring sentences. Complementary terms include words that commonly occur together like 'boy-girl' or 'play-fun'. Such terms may also reflect topic maintenance by referring to things that commonly occur together. The following sentence pairs reflect this type of semantic connection:

He fired the gun. A bullet grazed my ear.

The UFO landed. Aliens appeared.

Score 0 if complementary word pairs do not appear in neighboring sentences anywhere in the passage.

Scoring Companion

Reference			
Item	**Scoring Criteria**	**Example**	**Non-example**
1	**Every** 3rd person pronoun refers to a previously mentioned noun.	uses *he, she, they, it, him, his, etc.* the referent for every pronoun is clear	a pronoun with no referent, pronoun mismatch, or unclear referent disregard *I, you* and *it* used for setting
2	**Every** 3rd person pronoun refers to a noun or pronoun **in the previous or same sentence.**	*The boy was walking fast. He was headed home.* OR *The man was afraid so he ran.* OR *She walked. She laughed.* OR *She cried. Tears rolled down her cheeks.*	referent not in the previous sentence or clause *The girl ate some chips. The chips tasted very good. Then she drank her pop.* OR no pronouns used
3	Every noun used with a demonstrative has mention in the previous sentence.	uses *this, that, those, these, there,* or *here* before a noun or to refer to a noun *A dog came running at us. That beast was mean.* *I walked in the room. It was dark in there.*	*The dogs came running at us. We were scared. Those dogs were mean.* OR use of a demonstrative with or for a noun with no previous mention or no demonstratives used
4	'The' is used with a noun with an unambiguous referent previously mentioned or understood from world knowledge.	*the Prime Minister* *the manager* *The day was warm and sunny.* *A dog chased me up the street. I ran as fast as I could. The dog looked mean.*	'the' used with the first mention of a noun that is not a special case: *the cat, the officer* OR 'the' used with an unclear referent or never used with a referent from the text
5	**Every sentence** contains a reference to the **previous sentence**	reference to previous sentence through the correct use of *pronouns, demonstratives,* or *the definite article 'the'*	At least one sentence is not connected to the previous one by use of *pronouns, demonstratives,* or *the definite article 'the'*

Conjunction

Item	Scoring Criteria	Example	Non-example
6	Use of additive conjunctions to connect independent clauses	uses *and, also, or, another, in addition, additionally, as well as, furthermore, besides, nor* etc. *The children were playing and the boys were running.* OR *The dog was mean. And I was afraid.*	*The boy and girl were running.* *The ball was red and black.* *The children are laughing and playing.* OR additive conjunctions are not used
7	Use of 'when'	*When I got home I ate lunch.*	no use of 'when' OR 'When' is not used to join clauses. *"I'm coming over" she said. "When? I asked.*
8	Use of other subordinating temporal conjunctions	uses *before, after, until, while, as, since(time),* etc. to join clauses	no subordinating temporal conjunctions (besides 'when') used to join clauses
9	**Adverbs or adverbial phrases** used to mark shifts in time or sequence	*then, next, soon, first, second, third, next, last, finally, suddenly, later, all of a sudden, as soon as, the next day, later on, the following week, after that,* etc	no adverbs or phrases used marking shifts in time
10	Use of causal conjunctions to join sentences or clauses	uses *so, because, therefore, consequently, since (cause),* etc. *I was hungry. So I ate something.* OR *Since I was late, I hurried home.*	'so' used in a way that conveys degree rather than causation *I was so hungry, I could eat a horse.* OR conjunction does not connect two ideas *I went home because.* OR no causal conjunctions used
11	Use of adversative conjunctions	*but, however, although, yet, instead, except, though,* etc.	no adversative conjunctions used

Lexical Cohesion

Item	Scoring Criteria	Example	Non-example
12	Uses reiteration of a referent at least one time anywhere in the text (applies to nouns only)	superordinates *I saw a dog. The animal was huge.* OR synonyms *I saw a dog. The mutt was huge.* OR near-synonyms *I saw a dog. The beast was huge.* *note- these items will be signaled by the use of the definite article 'the' or a demonstrative.	repetition of a word *I saw a dog. The dog was huge.* OR no superordinates, synonyms, or near-synonyms are used
13	Use of complementary terms in **adjacent sentences**	*The gun fired. A shot rang out.* OR *We went to the beach. The sand was hot.*	such terms are not used in adjacent sentences or not at all.