
Theses and Dissertations

Spring 2010

Comparison of test directions for ability tests: impact on young English-language learner and non- ELL students

Joni Marie Lakin
University of Iowa

Copyright 2010 Joni Marie Lakin

This dissertation is available at Iowa Research Online: <http://ir.uiowa.edu/etd/536>

Recommended Citation

Lakin, Joni Marie. "Comparison of test directions for ability tests: impact on young English-language learner and non-ELL students." PhD (Doctor of Philosophy) thesis, University of Iowa, 2010. <http://ir.uiowa.edu/etd/536>.

Follow this and additional works at: <http://ir.uiowa.edu/etd>



Part of the [Educational Psychology Commons](#)

COMPARISON OF TEST DIRECTIONS FOR ABILITY TESTS: IMPACT ON
YOUNG ENGLISH-LANGUAGE LEARNER AND NON-ELL STUDENTS

by

Joni Marie Lakin

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations
in the Graduate College of
The University of Iowa

May 2010

Thesis Supervisor: Professor David F. Lohman

ABSTRACT

Ability tests play an important role in the assessment programs of many schools. However, the inferences about ability made from such tests presume that students understand the tasks they are attempting. Task familiarity can vary by student as well as by format. By design, nonverbal reasoning tests use formats that are intended to be novel. The popularity of nonverbal reasoning tests has increased substantially in recent years because of the increasing number of English-language learner (ELL) students in many U.S. school districts. Nonverbal tests are thought to eliminate the need for language in test items and to reduce cultural content. Formats on these tests are also assumed to be equally novel for all students. However, in at least one large study, researchers found substantial differences between the average performance of ELL and non-ELL Hispanic students on three of the most widely used nonverbal tests. Although these differences might reflect real variation in cognitive development, they may also reflect differences in knowledge of test formats and the testing practices used in U.S. schools.

In this study, I hypothesized that the score gaps between ELL and non-ELL students might, in part, be due to differences in test familiarity and that providing directions that include more practice and feedback might attenuate these differences. I drew from the research on universal design, dynamic assessment, and cross-cultural testing to develop three different types of directions with practice items. I then compared the effects of these three types of test directions on students completing a nonverbal figure analogies test. Figure analogies tests are generally among the best measures of reasoning abilities and are known to be quite difficult for young students. All directions were provided using video with English and Spanish audio and minor animations to concretize the instructions. The three types of directions were nonverbal-dynamic directions, verbal-dynamic directions, and a control condition that used standard test directions. The nonverbal-dynamic directions presented four practice problems that sampled the range of items on the test. Oral instructions and feedback were minimal. The verbal-dynamic directions presented the same four practice problems with more in-depth description and feedback. These

directions also described useful strategies for solving items. The standard test directions presented two sample problems with minimal instruction and feedback.

The sample consisted of 882 students in 40 first- and second-grade classrooms in 8 schools. A hierarchical linear model was used to control for similarity among students nested in classrooms and schools and to account for the assignment of treatment (type of directions) at the classroom level. The model included tests for main effects and interactions among treatment, ELL status, and grade. Results indicated that providing additional practice (the nonverbal-dynamic directions) led to small gains in performance, but that the more extensive set of directions (verbal-dynamic directions) were effective only for high-ability students. Contrary to the hypotheses, there was no interaction of ELL status with treatment. An unexpected finding was that use of teacher-read directions instead of video-based directions led to better performance for second-grade students. I conclude that test directions are an important means for improving test familiarity in young students, but that excessive standardization and lengthening of the directions may hinder performance. I also conclude that the choice of practice items and feedback are crucial considerations in the design of test directions.

Abstract Approved:

Thesis Supervisor

Title and Department

Date

COMPARISON OF TEST DIRECTIONS FOR ABILITY TESTS: IMPACT ON
YOUNG ENGLISH-LANGUAGE LEARNER AND NON-ELL STUDENTS

by
Joni Marie Lakin

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations
in the Graduate College of
The University of Iowa

May 2010

Thesis Supervisor: Professor David F. Lohman

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Joni Marie Lakin

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Psychological and Quantitative Foundations at the May 2010
graduation.

Thesis Committee: _____
David F. Lohman, Thesis Supervisor

Stephen B. Dunbar

Judith E. Liskin-Gasparro

Walter P. Vispoel

Catherine Welch

To my granddaddy, W. M. (Mac) Schneider, for always being proud of me.
And to my parents, for nature and nurture.

ACKNOWLEDGMENTS

I am sincerely grateful to my advisor, Dave Lohman, in particular for supporting this research study and in general for six years of patient mentoring. I am also grateful to my dissertation committee for their helpful feedback in developing this study.

I am also grateful to the many students, teachers, and administrators who participated in this study. In particular, I owe a great debt to Cindy G., Gifted and Talented Coordinator in the participating school district, for her incredible generosity of time and energy. And I am grateful to the two translators who assisted with this study—Antonio Heras and Angelique Dwyer—for sharing their expertise.

Several other individuals must be thanked for getting me through the Ph.D. process. Margaret Beier helped me get into graduate school in the first place and continues to be an invaluable mentor. Katrina Korb was always two steps ahead in the Ph.D. program, so I never had to figure out the bureaucracy on my own. Emily Lai and Jane Gressang have been my constant companions through the dissertation process. And, finally, Kathy Schuh has been an unofficial advisor for the past five years. I am eternally grateful to each of these people for their kindness and ready advice.

Finally, I am grateful to my family. None of this would have been possible without the support of my parents, sisters, and Daniel. Thanks for letting me do what I want to do.

ABSTRACT

Ability tests play an important role in the assessment programs of many schools. However, the inferences about ability made from such tests presume that students understand the tasks they are attempting. Task familiarity can vary by student as well as by format. By design, nonverbal reasoning tests use formats that are intended to be novel. The popularity of nonverbal reasoning tests has increased substantially in recent years because of the increasing number of English-language learner (ELL) students in many U.S. school districts. Nonverbal tests are thought to eliminate the need for language in test items and to reduce cultural content. Formats on these tests are also assumed to be equally novel for all students. However, in at least one large study, researchers found substantial differences between the average performance of ELL and non-ELL Hispanic students on three of the most widely used nonverbal tests. Although these differences might reflect real variation in cognitive development, they may also reflect differences in knowledge of test formats and the testing practices used in U.S. schools.

In this study, I hypothesized that the score gaps between ELL and non-ELL students might, in part, be due to differences in test familiarity and that providing directions that include more practice and feedback might attenuate these differences. I drew from the research on universal design, dynamic assessment, and cross-cultural testing to develop three different types of directions with practice items. I then compared the effects of these three types of test directions on students completing a nonverbal figure analogies test. Figure analogies tests are generally among the best measures of reasoning abilities and are known to be quite difficult for young students. All directions were provided using video with English and Spanish audio and minor animations to concretize the instructions. The three types of directions were nonverbal-dynamic directions, verbal-dynamic directions, and a control condition that used standard test directions. The nonverbal-dynamic directions presented four practice problems that

sampled the range of items on the test. Oral instructions and feedback were minimal. The verbal-dynamic directions presented the same four practice problems with more in-depth description and feedback. These directions also described useful strategies for solving items. The standard test directions presented two sample problems with minimal instruction and feedback.

The sample consisted of 882 students in 40 first- and second-grade classrooms in 8 schools. A hierarchical linear model was used to control for similarity among students nested in classrooms and schools and to account for the assignment of treatment (type of directions) at the classroom level. The model included tests for main effects and interactions among treatment, ELL status, and grade. Results indicated that providing additional practice (the nonverbal-dynamic directions) led to small gains in performance, but that the more extensive set of directions (verbal-dynamic directions) were effective only for high-ability students. Contrary to the hypotheses, there was no interaction of ELL status with treatment. An unexpected finding was that use of teacher-read directions instead of video-based directions led to better performance for second-grade students. I conclude that test directions are an important means for improving test familiarity in young students, but that excessive standardization and lengthening of the directions may hinder performance. I also conclude that the choice of practice items and feedback are crucial considerations in the design of test directions.

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	ix
CHAPTER	
I. INTRODUCTION	1
Ability Tests and Cultural Loading	3
Statement of the Problem	5
II. LITERATURE REVIEW	9
The Role of Test Directions	10
What is There to Learn About Tests?	13
Group Differences Between ELL and Non-ELL Students	16
Expert Advice on the Design of Good Directions	25
Universal Design Modifications Intended to Help ELL Students	42
Simplified Language Modifications	43
Limitations of Improving Directions	47
Conclusions About the Design of Test Directions to Improve the Validity of Tests with Novel Formats	50
Current Study	52
III. METHOD	56
Participants	56
Procedure	58
Materials	59
Design and Hypotheses	65
Data Entry and Missing Data	74
IV. RESULTS	78
Review and Results of Pilot Testing	78
Sample	80
Item analyses	80
Preliminary Information About Treatments	83
Multilevel Model	85
Differential Item and Bundle Functioning	95
Effects of Treatments by Quartiles	97
Matching Strategy	101
Teacher Comments	102
Summary	102
V. DISCUSSION	104
The Positive Effect of the Nonverbal Treatment	104
Narrow Effect of Verbal-Dynamic Directions	105
Negative Effect of DVD	106
Implications for Designing Test Directions	107

Limitations.....	112
Future Research.....	113
REFERENCES	114
APPENDIX A DIRECTIONS FOR ADMINISTERING STANDARD DIRECTIONS.....	126
APPENDIX B DIRECTIONS FOR ADMINISTERING VERBAL-DYNAMIC DIRECTIONS.....	134
APPENDIX C DIRECTIONS FOR ADMINISTERING NONVERBAL- DYNAMIC DIRECTIONS	146
APPENDIX D STUDENT TEST BOOKLET FOR STANDARD DIRECTIONS	155
APPENDIX E STUDENT TEST BOOKLET FOR VERBAL- OR NONVEBAL- DYNAMIC DIRECTIONS	166

LIST OF TABLES

Table	
2.1	Directions and examples from published tests30
3.1	Summary of sample55
3.2	Summary of item features.....58
3.3	Overview of research questions and model.69
3.4	Variables collected.....72
4.1	Classroom treatment assignment and DVD usage.....77
4.2	Descriptive Statistics.....78
4.3	Test characteristics by treatment and grade.....81
4.4	Test characteristics by treatment and ELL status.81
4.5	Coefficient estimates for each step of model building.89
4.6	Fit statistics.90
4.7	DIF and DBF results.....92

LIST OF FIGURES

Figure

2.1	WISC-IV Spanish Scale Scores by Percent Education in the U.S. (Data drawn from Weiss et al., 2006, Ch. 1)	24
3.1	Example of a figure analogy	57
4.1	Distribution of scores on FA test	79
4.2	Plot of percentile rank by FA score indicates sufficient floor and ceiling	79
4.3	Percent Rank by FA score defined within grade and treatment	80
4.4	Overall Effects in Full Model	91
4.5	Mean FA scores with standard error bars for quartiles defined within treatment and grade.....	94
4.6	Mean FA scores for quartiles defined with CogAT 6 Nonverbal SAS scores	94
4.7	Conditioning on CogAT 6 scores	96

CHAPTER I

INTRODUCTION

Culturally and linguistically diverse students perform less well than their peers on both ability and achievement tests (Jencks, 1998). These score gaps have become an urgent concern given the dramatic increase in the number of English-language learner (ELL) children in U.S. schools in recent years (Harris, Rapp, Martinez, & Plucker, 2007). These gaps are a major concern because of their magnitude (sometimes over 1 SD in size; Abedi, 2002) and because they persist for many years after ELL students have entered the U.S. school system (Ortiz & Ochoa, 2005).

One view of the gaps is that they reflect bias built into the school system. Students from low socioeconomic status (SES) and minority groups (which include many ELL students) are more likely to attend less effective schools and to have less qualified teachers (Thomas & Collier, 1997). In this view, the gaps in performance represent real differences in achievement or developed ability that must be addressed through instructional improvement. Another view is that the tests themselves cause at least part of the score gaps by drawing on culturally loaded content and item formats. Researchers have long argued that test features such as language-based directions or word-based math problems increase differences in performance between ELL and non-ELL students by introducing construct-irrelevant demands to the test items (Abedi & Lord, 2001).

Arguments about educational opportunity and cultural loading of test items are more often levied against ability tests than achievement tests because ability tests are assumed to penetrate the veneer of opportunity and provide a measure of a student's innate capacity for learning. Because of these assumptions, many believe that test bias alone must account for the lower performance of ELL and minority children. These critics question whether cognitive ability tests can yield valid inferences about the intelligence of these students without considerable modifications (Lewis, 2001).

Nonverbal ability tests have been suggested as a means of obtaining valid ability estimates for ELL students because nonverbal tests rely on figures and pictures rather than on language in the test items. Such item formats are thought to reduce the cultural load of the tests and therefore are expected to result in equal mean performance for ELL and non-ELL students (Ortiz & Ochoa, 2005; Powers & Barkan, 1986). However, researchers have found that even nonverbal test formats do not eliminate the differences in ability scores between ELL and non-ELL students, even when both groups of students come from the same cultural and socioeconomic background (Lohman, Korb, & Lakin, 2008). In fact, even after controlling for SES, Hispanic ELL and non-ELL children differed by approximately .5 to .6 SD on three popular nonverbal ability tests administered in one large study (Lohman et al., 2008).

Some researchers have suggested that the cause of persistent mean differences on nonverbal tests might be that while the items do not contain language, they still rely on language-based test directions to explain the task to students. For this reason, researchers such as McCallum, Bracken, and Wasserman (2001) argue that even the test directions should avoid language and rely on gestures and practice items to teach the task to the students.

However, the problem might be the exact opposite—perhaps nonverbal tests do not give *enough* explanation for students who are new to both the language and the school system to understand the task. Nonverbal tests are designed to be novel. Further, in an effort to reduce the effects of language, most use only short and simple instructions. This may be sufficient for some students, especially students who understand the language and who have developed strategies for approaching and solving such tasks. However, ELL students have both language and cultural barriers that may make it difficult for them to understand directions or to know how such test items are commonly solved. They are also less likely to have prior experiences with tests or test preparation (Pitoniak, Young, Martiniello, King, Buteux, & Ginsburgh, 2009). Perhaps these

examinees actually need more time to learn the rules of the task and to develop and hone a basic solution strategy before they attempt their first scored item. This is most efficiently accomplished with improved test directions that provide more opportunities to practice the task with feedback so examinees can be confident that they know the task demands and have a basic understanding of how to go about solving the test items.

Ability Tests and Cultural Loading

Ability tests play an important role in the assessment programs of many schools, especially for making placement decisions in programs for the talented and gifted (Callahan, 2005). A common misconception is that ability tests see through the effects of education and socioeconomic status to some innate intellectual capacity, but in reality, ability tests measure developed and well-practiced knowledge and reasoning skills (Ferguson, 1954; Thorndike, 1927). Rather than providing categorically distinct information from achievement tests, ability tests offer a different *perspective* on developed knowledge and skills that can be useful to teachers who want to adapt the pace and content of their instruction to students who differ widely in the speed at which they learn (Lohman & Hagen, 2002).

Generally stated, achievement tests are used to make inferences about how much a student knows with respect to a relatively well-defined domain and relative to criteria for proficiency that are usually based on grade-level standards. Ability tests, on the other hand, are used to make inferences about a more nebulous concept of aptitude in various domains. Because the domain is ill defined, such inferences can only be made relative to individuals with similar opportunities to learn (Anastasi, 1981). For achievement tests, the inference is “how much do they know?” and for ability tests, it is “how easily do they learn?” Thus, inferences about individual differences in efficiency of learning can be made only after controlling as much as possible for differences in motivation, opportunity to learn, or acculturation (Schwarz, 1971). Thus, the inferences made from ability tests

are more easily confounded by extraneous factors such as differences in how well students understand the task they are attempting.

Many researchers limit their discussion of the cultural information required for tests to the content of the items. Such researchers call for the use of nonverbal tests with figural stimuli with diverse groups of students, assuming the content will be equally unfamiliar to all. However, cultural background may also influence test performance through the familiarity of the item formats and particularly the types of reasoning that ability tests require. Researchers have found that young children perform better on tests that mimic the types of interactions that the children have with their caregivers. For example, Peña and Quinn (1997) found that children whose mothers engaged more frequently in labeling activities were more comfortable with vocabulary tasks, whereas children whose mothers did not engage in these activities were more comfortable with tasks involving information-seeking questions.

As Stanley (1971) observed: “Performance on many types of tests is likely to be in some measure a function of an individual’s ability to understand what he is supposed to do on the test” (p. 364). Differences in test familiarity as a result of test practice, even in culturally homogeneous populations, have long been known to impact scores on ability tests (Thorndike, 1922). To the extent that tests rely on relatively novel tasks that are not a regular part of the shared school curriculum, differences in task familiarity will play an important role in test performance.

Although differences in test familiarity can affect performance on any innovative item format, these differences are more likely on ability tests where novelty is more valued. Once students have been in school for several years, traditional item formats for reading and mathematics achievement tests are likely to be well practiced because these item formats are frequently encountered on classroom assessments. Additional practice on such items is unlikely to have a large impact on test performance. On the other hand, reasoning tasks, such as the analogies formats commonly used on nonverbal ability tests,

are rarely presented on classroom assessments. Such test formats are likely to show greater effects of differences in familiarity and are therefore more likely to be influenced by additional practice opportunities.

Thus, even though these nonverbal formats are intended to avoid cultural content, the cultural loading of item format may actually be exaggerated on nonverbal tests of reasoning. In such cases, test directions will play a crucial role in making valid score inferences about students who vary in their familiarity with the format prior to testing (Anastasi, 1981). Directions can bring students up to the same level of familiarity with the basics of the item format and help them to develop a basic solution strategy with which to begin answering items. Unlike other test preparation activities, test developers have more control over how test directions are presented, and they can assure that every student is presented with the directions prior to testing. However, the effectiveness of directions should not be assumed. In fact, directions should be scrutinized closely when there is a potential for some students not to gain as much from the directions as other students, such as when students vary in their cultural and linguistic background.

Statement of the Problem

Test directions may play a critical role in culturally heterogeneous schools through their role in equalizing test format familiarity. Nevertheless, test directions receive relatively little consideration in the research literature despite widespread agreement on their importance (AERA, APA, & NCME, 1999; Clemans, 1971).

Detterman and Andrist (1990) opine the following:

Unless instructions are systematically tested to ensure understanding, they may be a potent variable in explaining outcome. This would be particularly true of cross-cultural research and work with impaired populations, where there may be reasons to suspect differences in degree of understanding instructions and where mean differences between groups are theoretically important. (p. 389)

Although there has been some research on the impact of test directions on test performance, much of it has focused on college-aged students (Detterman & Andrist, 1990; LeFevre & Dixon, 1986; Whitely & Dawis, 1974). This is a serious flaw because by the time most students reach their teens, they have taken many different kinds of tests. This is especially true for students who have been admitted to college where acceptable scores on the SAT or ACT tests are necessary for admission. The finding that directions are disregarded (LeFevre & Dixon, 1986) or that practice and instruction have small effects (Detterman & Andrist, 1990; Whitely & Dawis, 1974) may not apply to younger students who do not have a long history of test-taking experience. In addition, the research on test directions has been particularly sparse with respect to populations with special needs such as ELL students.

I believe that the design of test directions can have a significant impact on the ability of examinees to engage meaningfully in a test task, particularly if they are unfamiliar with the task or have characteristics that make them more sensitive to the quality of directions provided. Directions could therefore create or exacerbate differences between groups. Specifically, I believe reduced-language directions do *not* convey the demands of novel tasks well to students because they do not offer specific strategies for solving items and thereby increase the difficulty of inferring a procedure for answering items. I believe the use of more language in directions is important and that supporting language comprehension with dynamic visuals and video will result in more valid test directions than attempting to remove language from the directions and relying on practice items to convey the task.

In this study, I compared the effects of three types of directions on the performance of ELL and non-ELL students completing a nonverbal reasoning task. The sample included young students (first and second grade) in schools with large ELL populations. The control treatment involved standard directions with two practice items and basic descriptions of the figure analogies task. These directions were based on the

directions used by existing tests. The control treatment was compared to nonverbal-dynamic directions that provided four practice items with minimal language and to verbal-dynamic directions that expanded the practice provided by the nonverbal-dynamic directions by adding opportunities for teacher-led discussion and strategy training. The training provided in the verbal-dynamic directions promoted strategies that are often used by more able problem solvers. These strategies included a systematic/analytic strategy where students construct their answers mentally before looking at the answer choices and a verbalization strategy where students put the puzzle rule into words before considering the answer choices.

Critics of previous work on test directions have argued that if a short set of directions could eliminate score gaps between cultural groups, it would mean that the test itself must not be a good measure of ability and the results would not generalize to other tests or educational outcomes (Humphreys, 1976). It may in fact be too much to hope that five minutes of quality instruction could help students develop a systematic approach to problem solving that might have been taught since early childhood in middle- to upper-SES students. However, the very premise of using test directions is that a short amount of instruction will adequately familiarize students with the task and allow them to do their best. If such a premise is legitimate, then the quality of test directions will determine their effectiveness in equalizing familiarity. As such, the measurement field would benefit from identifying characteristics that distinguish effective from ineffective directions. Furthermore, given that test directions have changed very little for during the last half-century despite considerable shifts in the demographics of the U.S. population, a reevaluation of the adequacy of directions within the growing population of ELL students would provide important validity evidence for tests.

My observation from working with young students has been that many of the bright students already have strategies relevant to the task, but initially do not apply these strategies and instead respond impulsively. For these students, a small amount of

prodding or encouragement to slow down and think was enough to change the way they approached other items. I believe that if my directions are successful in promoting a thoughtful approach to the task, they will improve the performance of bright students who come from families that do not spend time on test- or school-preparation activities outside of school. Further, if format familiarity is an important contributor to score gaps, I expect ELL students who fall into this category to be especially benefitted.

CHAPTER II

LITERATURE REVIEW

Test directions are a crucial but often overlooked component of test development. Test directions serve to explain unfamiliar tasks to test-takers and are often the only means that test developers have of diminishing preexisting group differences in test familiarity. If groups differ in their understanding of the test task *after* the directions have been completed, bias can be introduced to the test scores, resulting in greater score differences between the two groups that do not reflect real group differences in ability.

Test formats differ in their intended level of novelty and in the frequency with which they are used in classroom activities. Test formats that are novel and infrequently used in classrooms are more sensitive to differences in familiarity and advantage those who have access to practice opportunities. Preexisting differences in familiarity are further exacerbated when some test-takers are unable to comprehend the test directions. This is often true for students whose first language is not English but who are students in an English-speaking school.

In this study, I evaluated the effect that test directions have on ability test scores for groups of students who were and were not English-language learners (ELL). The test directions studied varied in the amount of training and practice they provided. I expected to find that directions with minimal explanations were inadequate for teaching novel test tasks, decreasing scores for both ELL and non-ELL students compared to enhanced directions with more practice and feedback. I also expected to find that directions that used language combined with dynamic video demonstrations would result in higher test performance for English-language learners and narrow the gap between them and their native-English-speaking classmates relative to standard directions.

My goal was to develop a single set of test directions that helped both ELL and non-ELL students perform their best on a figure analogies task. The purpose was to

explore the important role that test directions play as an instructional activity aimed at helping examinees develop complex problem-solving skills required for the test. Directions that do not provide adequate instruction demand greater levels of inference from the examinee and add additional burdens to their comprehension of the individual test items. When examinees differ in their familiarity with tests *or* when directions are differentially effective, the increased cognitive load is not evenly distributed among examinees. Differences in cognitive load may introduce bias into the test scores.

In this chapter, I review several broad issues related to my topic. First, I review the role of test directions in teaching the task and reducing unfamiliarity in examinees. Second, I review evidence on the differences between ELL and non-ELL students that may influence test familiarity including socioeconomic status, culture, and exposure to testing as it is practiced in the U.S. Third, I review the guidance provided by testing experts on how to design valid test directions. Most of the useful guidance comes from older work, but sheds light on how directions can be made effective for diverse examinee populations. Fourth, I review recent advice on Universal Design that addresses how to create tests that minimize the need for accommodations for students with special needs (Thompson, Thurlow, & Malouf, 2004). Finally, I summarize what I perceive to be the key features of high quality test directions that benefit both ELL and non-ELL students.

The Role of Test Directions

Test directions are an instructional tool whose purpose is to teach examinees the sometimes complex task of how to solve a series of test items. Most directions rely heavily on examples accompanied by oral or written explanations of the task. The effectiveness of a set of directions depends not only on the characteristics of those directions, but also on the characteristics of the examinees, including their language proficiency and familiarity with the test tasks.

From the perspective of measurement professionals, test directions are important because they familiarize examinees with the basic rules and process of answering test items. The assumption is that there are certain rules of the game that examinees must know to answer items correctly and efficiently. If test items are answered as intended by the test creator, the resulting scores are more likely to provide a valid indication of the intended construct (Clemans, 1971). If examinees solve items incorrectly simply because they did not understand the format of the item or what they were supposed to do, then their responses usually cannot be considered a valid indicator of ability.

On some tests, such as classroom tests, traditional achievement tests, and even ability tests for college applicants, it is assumed that understanding the task is trivial and so the directions are perfunctory (Scarr, 1981). The directions are short because the examinees have been exposed to many similar tests (e.g., vocabulary or mathematical computation tests) or the examinees are assumed to have engaged in test preparation activities. Some researchers even complain that examinees entirely ignore the test directions and skip straight to the test items as soon as they are able (James, 1953). In these cases, low-quality directions are expected to have a negligible impact on test scores.

On other tests, the formats are less likely to be familiar. This is especially true for cognitive ability tests designed for younger schoolchildren. This is because ability tests in particular include novel tasks that challenge students' analytical skills. This is a legacy dating back at least to Binet, who constructed his test items to be equally unfamiliar to children with and without extensive prior education (Binet & Simon, 1908/1961). Stern (1914) praised Binet and Simon's method and argued for intelligence tasks that were more novel and uninfluenced by education or a privileged home life. Today, ability tests continue to use a variety of novel formats that differ from familiar classroom activities and assessments.

It has long been acknowledged that practice can significantly affect performance on novel tasks and that differences in familiarity can lead to higher scores (Thorndike,

1922). Several researchers comparing the effects of training on item performance found that more complex and unfamiliar formats showed the greatest gains in response to training (Bergman, 1980; Evans & Pike, 1973; Glutting & McDermott, 1989). The use of novel formats means that practice effects are greater even without explicit training or feedback. For example, Thorndike and Hagen (1974) found that the practice effects from retesting students on the CogAT with counter-balanced parallel forms were three times greater for the Nonverbal Battery (.22SD) than the Verbal Battery (.07SD), which has much more familiar formats than the Nonverbal Battery.

From their everyday school experiences, students learn how to use the information in a mathematics word problem to arrive at an answer or how to generate an answer to a vocabulary question. They may not know the content required to answer correctly, but, in general, they have assembled strategies for these types of questions. This is not always true for ability tests and their novel test formats. In this case, even if the test-takers know generally what the task is (e.g., find an analogical solution) and they understand the examples given, they still may not have a relevant strategy assembled for constructing an appropriate answer on their own. In such a case, examinees must first assemble a preliminary strategy and monitor its effectiveness based on any feedback they receive during the practice items. This aspect of developing an appropriate strategy is where directions, examples, and feedback play a crucial role for examinees unfamiliar with the format. When examples are too simple or too few and do not adequately represent the problems on the test, examinees may not generate an appropriately nuanced test-taking strategy (Klausmeier & Feldman, 1975).

Overly simple examples and directions may lead to faulty inferences for some examinees who are not necessarily the least able students. For example, on Form 6 of the Cognitive Abilities Test (CogAT), it was observed on the Verbal Classification subtest that some students erroneously assumed that they were looking for superordinate terms in an item rather than for another member of the category (Lohman & Al-Mahrzi, 2003). So

for an item stem “blue, red, green”, the examinees would select “color” as the answer rather than the intended answer “brown”. Even though the practice item and directions clearly specified that the examinees needed to find another member of the category, an erroneous strategy was unintentionally reinforced by having that type of superordinate distracter on many of the early items on the test. It is clear that test directions, examples, and early test items must work together to prevent misunderstandings. Without a clear explanation of what the task is and careful scaffolding of understanding, students can persist with an ineffective solution strategy and receive test scores that do not reflect their true ability level.

What is There to Learn About Tests?

We know from extensive work on practice effects (Kulik, Kulik, & Bangert, 1984; Thorndike, 1922) that examinees improve their test scores when repeatedly administered similar tests. What is the examinee learning that results in practice effects on later examinations even when the items are not the same? Four components of problem solving stand out: (1) learning to attend to important elements; (2) assembling a working strategy; (3) learning the range of permissible solution rules; and (4) basic test-wiseness and test-taking skills.

First, the examinee is learning which task features are important. The ability to select relevant information from a field of competing salient features is an important source of variability in task performance (Markman & Gentner, 1993; Stenning, & Monaghan, 2004). Detterman and Andrist (1990) expressed surprise that college-aged examinees sometimes could not figure out a very simple (from their perspective) task without directions. “Evidently, an important part of performance on even very elementary cognitive tasks is the development of an understanding of what is expected on the task, or what might be called a mental model... of task performance” (Detterman & Andrist, 1990, p. 388). They concluded that tasks are often not as simple as they appear. Indeed,

the researchers who have developed and tested the items seem likely to underestimate how inscrutable their tasks can be to test-naïve examinees.

The second benefit that examinees gain from a practice test is the development of a set of workable strategies to apply when attempting to solve particular item types. For example, previous research has shown that putting a verbal analogy into the form of a sentence can improve performance and that this strategy is common among proficient analogy solvers (Bridgeman & Buttram, 1975; Sternberg & Nigro, 1980). Having created such a strategy from one exposure to analogies items, examinees would be more efficient in solving analogies when they were encountered on a later test. As another example, on speeded tests, examinees perform better if they make accurate judgments of how quickly they must work. Previous experience with a test may give an examinee direct experience with how quickly to work. In general, strategies are essential to strong test performance and are often learned through repeated exposures to a test. When such strategies have been developed by one group and not another, they can be a source of construct-irrelevant variance in scores.

Third, the examinee learns the range of solutions that are acceptable on a test. Sometimes these limits are explicitly stated in the directions, as on number series items, where the examinees are told that only basic arithmetic operations are used in the patterns to be discovered. At other times, these limitations are not stated and the examinees must discover them on their own. For example, good solutions to analogies are often based on implicit rules that would be difficult to describe to examinees without previous testing experience. On figural analogies items, students are often asked to infer the rule from multiple examples or given the (mostly unhelpful) guide of saying to themselves, “A is to B as C is to what?” It is difficult to specify for a child what it means to discover a relationship between A and B that can be transferred to a sometimes quite distant C term.

Because the analogy format is abstract and formal, Piaget was doubtful that young children were capable of this type of reasoning (Piaget, Montangero, & Billeter, 1977, as

cited by Goswami, 1991). In her review of research on the development of analogical reasoning, Goswami (1991) questioned whether young children failed to solve analogies because they had not yet developed the capacity for analogical reasoning or if they lacked knowledge of what constituted a good answer in an analogy. Other item formats have the same challenges: Stenning and Monaghan (2004) argued that unfamiliarity with formal logic in novice problem-solvers was an important source of misunderstanding in logical reasoning studies. Extensive experience with item formats may be required for examinees to learn the full range of implicit task rules.

Finally, students become more comfortable with the process of testing when they have positive previous testing experiences (Anastasi, 1981; Ortan, 1960) and have developed positive expectancies for performance (Budoff, Gimon, & Corman, 1974). Glutting & McDermott (1989) showed that this was true for both preschoolers and older examinees because examinee anxiety was greater when facing a new and unknown test. Part of this confidence in the testing situation comes from test-wiseness, which includes basic test-taking skills like “following group directions, managing time wisely, maintaining a sustained effort in a structured situation, eliminating unreasonable answer choices, utilizing test formats, using key words to locate the answer, guessing, and substituting answer choices” (Crowe, 1982, abstract). A more positive testing experience occurs when examinees know exactly what to anticipate on a test and know that they have general strategies for answering items.

As James (1953) advocated, the goal “is to give all candidates the minimum amount of coaching which will enable them to do themselves full justice on the test. All candidates are then given a flying start and not just some of them” (p. 159). Trying to compare two students—one of whom has practiced the task many times and the other student who is seeing it for the first time—will result in invalid comparisons. Differences on any of these dimensions of test knowledge can lead to differences in test scores. Therefore, test directions are important because they are the last opportunity to level the

playing field on all four dimensions of test familiarity. If directions effectively instruct all the students on the format, then the students will have a better opportunity to show what they know on the test items (Ortar, 1972). Assuming the test is not speeded, the extra practice, even if administered to both test-wise and test-naïve students, should only level the playing field. The test-savvy student is not expected to make gains as a result of better test directions.

Group Differences Between ELL and Non-ELL Students

When differences in examinee background knowledge are ignored, the reasoning process and strategies used can vary across examinees. Ignoring variation in the underlying processes leads test users to assume too much commonality in the test construct across examinees (Leighton, 2004; Stenning & Monaghan, 2004). Hessels and Hamers (1993) further argue that when cultural differences are not considered, test scores are interpreted with the false assumption that “children taking tests understand directions, consider all possible responses before choosing the correct one, concentrate on one item at a time, are not distracted by other items, and are involved and attentive during the entire test” (p. 289). They argue that, in fact, group differences in familiarity with testing impacts each of these characteristics to an unknown degree.

Hispanic ELL students have a number of social and cultural differences—in addition to the linguistic differences— that may contribute to the widely documented differences in test performance. I am interested specifically in the differences that may lead to gaps in test familiarity and that might be eliminated by better-designed directions.

Socioeconomic Differences

Socioeconomic status (SES) is known to have a powerful effect on achievement through both educational opportunity and parenting practices including the quality of linguistic experiences of children (Hart & Risley, 1995). According to an AFT Policy Brief (2004), Hispanic students— and especially ELL Hispanic students— were more

likely than Black or White students to come from families that make less than \$30,000 per year. Hispanic students were twice as likely as Black students to have parents who completed less than a high school diploma level of education.

Coming from a migrant background presents even more risk factors as children from migrant families are more likely to move during the school year, disrupting their education through frequent adjustments to new school systems (AFT Policy Brief, 2004). In a 2002 survey, Hispanic students were found to be much more likely to come from migrant families than other students. Seven percent of Hispanic students came from migrant families, with Hispanic students making up 80 percent of the migrant student population (AFT Policy Brief, 2004).

Coming from low SES and especially migrant background can influence test familiarity in a variety of ways. First, engagement in the school community fosters familiarity with what tests are used and how the scores are used to make instructional decisions. Such knowledge can influence motivation to perform well on the test. Second, moving schools frequently can lead to missed opportunities for test preparation appropriate to the tests that the student actually completes. Students with greater access to test preparation or the mainstream culture are more likely to have the knowledge that will lead to accurate task comprehension and effective strategies for reaching the desired answers (Leighton, 2004).

Carman and Taylor (2009) found that students from low SES backgrounds received lower scores on the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1996) compared to students from middle- to high-SES backgrounds, despite that test's use of a nonverbal figure matrices format. Other studies have confirmed that SES may contribute to test score gaps on matrices tests through differences in test familiarity. For example, Budoff et al. (1974) developed group-administered learning potential tests for Raven's Progressive Matrices. The learning potential tests followed a pretest-training-posttest procedure. In the training activities, students completed items by drawing their answers

and received feedback from the teacher. Practice items began with familiar patterns like the U.S. flag and built up to analogies that were more complex. Budoff et al.'s motivation for developing the learning potential activity was a belief that ELL students "differ in familiarity and experience with particular tasks, have a negative expectancy of success in test-taking, and are less effective in spontaneously developing strategies appropriate to solving the often strange problems on a test" (p. 235). Through practice and training on the tasks, they sought to minimize these differences and to give those students an experience of success prior to the posttest. In a sample of 188 Spanish-speaking students ages 6-14, Budoff et al. confirmed that their learning potential posttest correlated better with achievement for ELL students compared to their pretest scores. They also found that indicators of SES and English proficiency were correlated with pre-test but not post-test performance.

Cultural Differences in Child-Rearing Practices, Family Values, and Understanding of the Purposes of Tests

Ethnographic and other observational research has long been used to document differences in child-rearing practices between social class groups. In addition to SES differences, there are apparent cultural differences between mainstream middle-class White families and Hispanic, Latino, and Chicano families with respect to their parenting behaviors, family values, and beliefs about testing purposes. For example, Heath's (1983) widely cited study of minority and low-SES Black students showed how home life and early social interactions influence children's familiarity with certain question-answer norms, ways of approaching tasks, and time limits. These differences sometimes caused conflict in the classroom with White, high-SES teachers. Heath concluded that the teachers' familiarity with urban, middle-class styles of play and questioning led to better rapport and fewer behavioral complaints with students who shared their own background.

Heath's study exemplifies the pervasive effects such cultural differences can have on the educational experiences of children.

More directly relevant to this literature review is research comparing the use of language in the parenting of Hispanic, Latino, and Chicano mothers. For example, Peña and Quinn (1997) found that Chicano mothers were more directing with children and used modeling and visual cues frequently. Puerto Rican mothers used more nonverbal directions with their babies. Research summarized by Peña and Quinn indicated that White, middle-class mothers used more labeling than Puerto Rican mothers, resulting in differences in test performance. Peña and Quinn related these differences to observations of Puerto Rican children working on picture labeling (vocabulary) tests. They observed that the children had more difficulty understanding these vocabulary tasks than they did understanding familiar information-seeking interactions that elicited descriptions, functions, and explanations.

In a similar study, Greenfield, Quiroz, and Raeff (2000) reported that "Mexican immigrant parents used questioning as a conversational strategy less than did first-generation Mexican American parents, who were more educated and assimilated to US culture" (p. 1120). They concluded that these differences reflected acculturation of the first-generation parents and led to better educational outcomes of their children. Interestingly, Greenfield et al. also found that Hispanic parents in the sample often reported beliefs that learning occurred through observing others and imitating them without explicit instruction. These beliefs and practices likely affect how children are prepared for standardized tests.

Other research confirms differences in the use of language in the home. Brooks-Gunn and Markman (2005) compared the parenting styles of White, Black, and Hispanic mothers in the context of the children's school readiness. They found that the reading behaviors differed significantly between White and Hispanic mothers with the latter reading less frequently to young children and owning a narrower range of books and

other reading materials in the home. They concluded that these differences, which are present in both English- and Spanish-speaking Hispanic families, affected the school readiness of the children.

Unlike parenting behaviors, family values for education may not differ that much across cultural groups, or, to the extent that they differ, they may not affect school success. Greenfield et al. (2000) surveyed the theories of other researchers who claimed that agrarian values held by many Hispanic families might be a detriment to their children's school achievement. In contrast to the predictions of these researchers, Greenfield et al. found mixed and sometimes beneficial interactions between the beliefs of families with agrarian values and schools with academic-occupational values. Overall, Hispanic parents interviewed by Greenfield et al. and Reese, Balzano, Gallimore, and Goldenberg (1995) were found to value education greatly regardless of whether their home environment was actually well-suited to supporting such goals. Thus, Hispanic families appear to share a value for education and a belief that education leads to a better life outcomes.

Despite these similarities, family values and practices appear to influence beliefs about the purposes and value of testing. Compared to White parents, Hispanic parents were less likely to see tests as benefitting their students and less likely to understand the role tests played in education (Pitoniak et al., 2009; Solano-Flores, 2008). Similarly, students new to the U.S. school system differed in their understanding of the purposes of testing and the basic strategies of completing tests (Reese et al., 1995). They also differed in their understanding of how to access test preparation materials and overestimated how much improvement could be expected from re-testing (Walpole et al., 2005). Such differences are believed to impact students' motivation to perform well on these tasks and to exert effort (Schwarz, 1971).

Monroe (as cited by Kopriva, 2008) found that many parents and students new to the U.S. did not understand how tests are administered and used in the U.S., and

sometimes had negative evaluations of the utility or fairness of tests, or had greater worries about the importance of tests in determining educational opportunity. Similar results have been found in high school students as well (Walpole et al., 2005). In some cases, parents who did not grow up in the U.S. were not aware of the availability of services for gifted students or what was required to obtain those services (Harris et al., 2007). When the tests have high stakes (such as for gifted and talented identification or college admissions), there are often large differences between White and minority students in the perceived importance of preparing at home for school-administered tests (Briggs, 2001; Chrispeels & Rivero, 2001; Kopriva, 2008; Solano-Flores, 2008; Walpole et al., 2005). Maspons and Llabre (1985) found that college-aged Hispanic students generally lacked test-taking skills necessary for U.S. testing systems and that over half of the students surveyed were unfamiliar with multiple-choice items. Likewise, Dreisbach and Keogh (1982) found that test-wiseness training, devoid of test-relevant content, improved the performance of Spanish-speaking children from low SES backgrounds.

Despite differences in preparation for testing, some research shows higher rates of compliance among Hispanic students completing individually administered tests. For example, Glutting, Oakland, and Konold (1994) found that test-session behaviors (compliance, agreeableness, and effort) was similar or better for Hispanic students compared to White students. Oakland and Harris (2009) found similar results in a sample of ELL Hispanic students. These researchers conclude that test-session behaviors are related to cognitive abilities, but are otherwise not related to cultural background. On the other hand, Frisby (1999) found small effects indicating that raters scored Hispanic examinees higher on undesirable test session behaviors (compared to White and African-American students) after controlling for ability, though this study was limited by sample size.

Differences in Sensitivity to Test Quality

ELL students are more sensitive to some aspects of test design. For example, ELL students were distracted more often (17%) by construct-irrelevant text and images than their nondisabled, English proficient peers (8%; Johnstone, Thompson, Bottsford-Miller, & Thurlow, 2008). When images are clear and relevant, they can also differentially aid ELL examinees. Martiniello (2003, 2008) also found that ELL students self-reported using the graphics on a test item to make sense of test questions and support language comprehension. Scarr (1978, 1981) also suggested that minority children (ethnic or linguistic) might gain the most from better test directions. In interviewing inner city Black high school students, she asked them if the Raven's Progressive Matrices test made sense to them and whether the directions helped them understand the task. She found that they generally were confused and frustrated by the format and directions. She then worked with them to develop a better set of directions to explain to students from their background what the task is and what they should do.

Score Gaps

There is abundant evidence that ELL students have lower mean scores on both achievement and ability tests when compared to White or non-ELL Hispanic students. On a state achievement test, Abedi (2002) reported mean differences of 1.0 and 1.4 SD for reading achievement in grades 2 and 9, respectively. For math achievement, differences were smaller, .61 and .88 SD, respectively. On the 2004 long-term NAEP, the score differences between students classified as ELL and non-ELL ranged from .56 to 1.1 SDs for the math test and .82-1.1 for reading when tests were administered with appropriate test accommodations, which did not include translations (National Assessment Governing Board, n.d.). For the 2007 NAEP mathematics test, mean differences for Hispanic and White populations ranged from 0.68 to 0.79 SD (across grades 4-12) whereas differences for ELL and non-ELL populations ranged from 0.88-

1.07 SD. Likewise, for the 2007 NAEP reading test, mean differences between Hispanic and White students ranged from 0.58 to 0.76 SD whereas differences for ELL and non-ELL populations ranged from 1.01-1.20 SD.

Ability tests often show mean differences as large as or larger than achievement tests. In a study using both ability and achievement measures, Palmer, Olivarez, Willson, and Fordyce (1989) found mean differences of 0.8SD between the Hispanic non-ELL and Hispanic ELL students on the WISC-R performance composite, 1.5SD on the WISC-R verbal composite, and 1.24SD on the K-ABC achievement total (effect sizes calculated by the author based on the Palmer et al., 1989, data). Even when only comparing low-SES Hispanic ELL and non-ELL students, differences are large. Lohman et al. (2008) found differences in test performance on nonverbal tests—Raven’s Progressive Matrices, Naglieri Nonverbal Ability Test and the Nonverbal Battery of the CogAT—with magnitudes of 0.4 to 0.6 SD for the ELL vs. non-ELL comparisons. Since the test items themselves are nonverbal, figural stimuli, it is possible that some portion of the difference reflected difficulties in understanding the test directions and the simplistic practice items provided there. Note that when appropriate, directions were given in Spanish, so it was not purely a language issue. As both of these studies demonstrate, nonverbal tests often show surprisingly large mean differences despite their reputation for diminishing mean differences between ELL and non-ELL students.

In another surprising result, controlling for social class differences and ethnicity did not eliminate differences on tests in the recent standardization of the Spanish form of the Wechsler Intelligence Scale for Children (WISC-IV Spanish; Wechsler, 2005; Weiss, Saklofske, Prifitera, & Holdnack, 2006). The target population for this adaptation of the WISC-IV was Spanish-speaking bilingual children who had been in the U.S. schools for no more than 5 years. The subtests that require knowledge of English were all carefully translated or adapted to Spanish. One analysis examined the effect of U.S. schooling on the verbal comprehension and perceptual reasoning scores of ELL students. The

researchers expected that participation in U.S. schools would have a large effect on verbal scores and little or no effect on the nonverbal/perceptual reasoning scores. However, the opposite effect was observed—children who had received more of their education in the U.S. schools performed markedly better on the perceptual reasoning index (about 10 IQ points) but only slightly better on the verbal comprehension index (about 2 IQ points). See Figure 2.1.

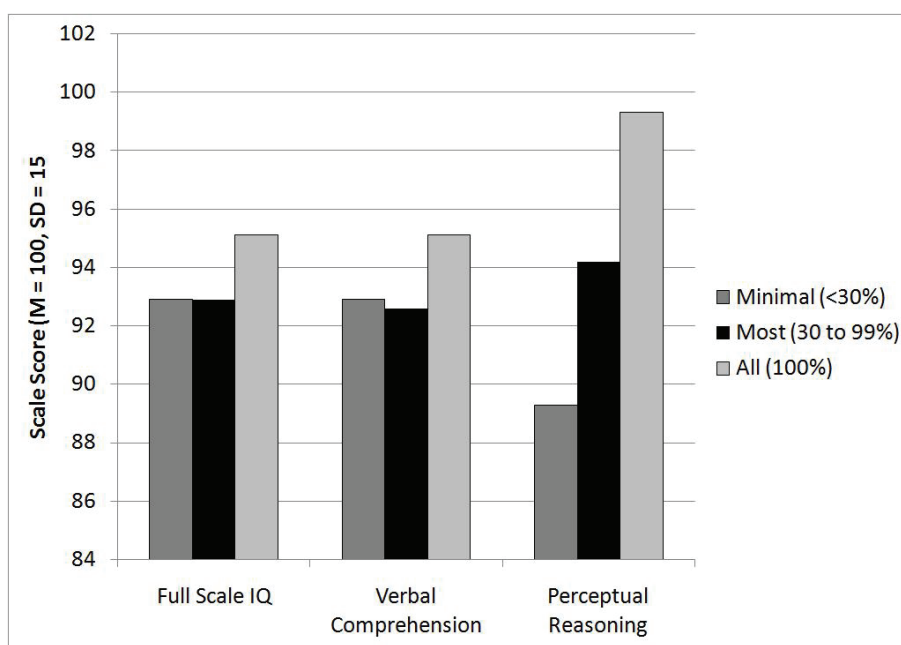


Figure 2.1. WISC-IV Spanish Scale Scores by Percent Education in the U.S. (Data drawn from Weiss et al., 2006, Ch. 1).

The substantial differences between ELL and non-ELL students on a wide range of nonverbal reasoning tests suggest that, at least as presently administered, these tests do not eliminate the effects of education and, in fact, may even enhance their effects. Perhaps the transition to a new culture and school system causes a general delay in cognitive development. A second, more optimistic, explanation is that ELL students simply have less experience than other children with the kinds of tests and test-like tasks

used on nonverbal reasoning tests. If this is the case, then improving test directions to help students unfamiliar with the format may have a significant impact on the test performance of ELL students.

Correlational Data

Ability tests show somewhat lower correlations with other tests for ELL students (Cathers-Schiffman & Thompson, 2007). However, these differences are not apparent when regression slopes are considered (Lakin & Lohman, 2009; Palmer et al., 1989). Lakin and Lohman (2009) found that reduced variability in the scores of ELL students (which affects correlations but not regression slopes) could account for the difference in correlations. Differences in variability could be due to lower reliability of the tests (Lakin & Lai, 2010) or to greater proportions of ELL students scoring at chance levels (Lohman et al., 2008). Poor performance on nonverbal tests cannot be due directly to the content of the items, but could be due to misunderstanding the directions or differences in the use of systematic analytical strategies.

Expert Advice on the Design of Good Directions

Several measurement texts tout the importance of good directions and offer some advice on the features of good directions (e.g., Clemans, 1971; Cronbach, 1984; Traxler, 1951). However, specific guidance for how to create directions is less common. To some extent, this is because the best directions differ by test and item types. However, the research also lacks systematic efforts to discover exactly what features make for good directions for a given format. The attention to the construction of test directions has also diminished over time, as exhibited by the content coverage of successive editions of *Educational Measurement*: Lindquist (1951), Thorndike (1971), Linn (1989), and Brennan (2006). The first two editions each contained a full chapter on test administration that contained extensive advice on creating the test directions (Clemans, 1971; Traxler, 1951), whereas the latter two editions restricted their discussion to issues

of standardization with little specific advice on developing test directions (Bond, 1989; Cohen & Wollack, 2006; Millman & Greene, 1989).

Given the infrequent inclusion of guidance on the development of directions in recent measurement texts, test directions may appear easy to construct. However, when actually trying to write directions, it can be quite difficult to explain an unfamiliar task clearly and succinctly. This is especially true when writing directions for children for novel item formats since even basic concepts must be minimized (Kaufman, 1978). Children have less experience with tests and need simple language and adequate practice.

From my work on the CogAT, a review of the literature (especially Cronbach, 1984), and examining ability similar tests, I have compiled a list of guidelines for creating directions and practice items. Some of these rules include:

1. “Betty Hagen rule”: Do not start the test until you are sure every child understands what he or she is supposed to do.
2. Use the shortest directions possible with the simplest language (Abedi, 2006).
3. Provide visual information to support verbal instructions as much as possible (Mayer, 2001).
4. Try to introduce a basic strategy for solving the early items so that students feel successful and can adapt the strategy as items increase in difficulty (Budoff et al., 1974).
5. Use more rather than fewer practice items, though how many items and how they are used varies across tests and examinees (Anderson, Fincham, & Douglass, 1997).
6. Provide feedback on the answers examinees give (Shute, 2007; Thorndike, 1927).
7. The first few test items (“teaching items”) should start easy and increase in difficulty so as to lead the child to develop appropriate strategies by gradually introducing more complex rules (Cronbach, 1984).

8. The first few items should not provide opportunities to succeed using inappropriate strategies (see earlier discussion of superordinate terms).

Clemans (1971) provided a compelling argument for the importance of directions to test validity and outlined the proper procedure for creating test directions:

The precision of a measuring instrument is a function of the clarity of the directions for test administration; these must have exactly the same meaning for the author, the examiner, and the examinee. (p. 189)

Clemans also adds:

The test author must keep in mind that the items are only part of the stimulus situation and that the directions are always a critical component of it. It is the combination of the two that provides the total stimulus and the basis for measurement. (p. 190)

Modern sources underscore the important role of test directions. According to the Joint Standards, the goal of test directions is to provide “sufficient detail so that test takers can respond to the task in the manner that the test developer intended” (AERA et al., 1999, p. 47). The guidelines from “Universal Design Applied to Large Scale Assessments” (Thompson, Johnstone, & Thurlow, 2002) state that assessment instructions should be easy to understand, regardless of a student’s experience, knowledge, language skills, or current concentration level. Directions and questions need to be in simple, clear, and understandable language for all examinees.

Cross-Cultural Research

The recommendations made from a cross-cultural perspective can also be useful when designing test directions for the heterogeneous U.S. school population. van de Vijver and Poortinga (1992) have worked extensively on the issue of cross-cultural testing and offer several helpful suggestions. First, they suggest that the directions should be elaborate and should contain a sufficient number of examples to convey the task rules. This advice stands in contrast to the minimal test directions often provided. Second, the directions should contain exercises for the examiner to ensure that the examinee has

understood the directions. This is rarely a component of traditional test directions. Third, the items should “incite the intended action” (p. 20), meaning that, to the extent possible, the items should be self-explanatory, encouraging the correct approach or possibly so similar to common school tasks that the task demands are easily recognized. Finally, the directions should “rely minimally on explanations in which the correct understanding of the verbalization of a key idea is essential” (p. 20). Although some might interpret this to mean favoring nonverbal explanations, I believe they meant that verbal directions should be supported by nonverbal clues (pictures, diagrams) in the directions, use basic concepts and language whenever possible, and rely on examples to help examinees understand the intended meaning. Indeed, van de Vijver and Poortinga (1992) argue that *increasing* the amount of instructions and practice can actually *reduce* the cultural loading of a test rather than increasing it. More language decreases how critical it is that students quickly grasp the rules of the task from a few examples. A liberal number of examples and exercises, they say, can overcome relatively small group differences in familiarity.

Biesheuvel (1972) likewise argued that familiarization with test demands is essential before beginning a test in a cross-cultural context. Multiple approaches—elaborate demonstrations, use of vernacular or mime, ensuring motivation—may be necessary. Practitioners often fail “to recognize the desirability of prolonged pre-test practice in dealing with the material which constitutes the content of the test and of *introducing a learning element in the test itself*” (Biesheuvel, 1972, p. 48, emphasis mine).

How Do Popular Intelligence Tests Explain the Task?

To gather more information on the state of the art in test directions, I reviewed several popular intelligence tests to determine how they presented the task directions and examples. Specifically, I compared how tests that used a standard set of directions (as opposed to dynamic directions) explained figure or matrix analogy problems similar to

those used in this study. The tests I reviewed were the Cognitive Abilities Test (CogAT; Lohman & Hagen, 2001), Comprehensive Test of Nonverbal Intelligence (cTONI; Hammill, Pearson, & Wiederholdt, 1996), Kaufman Assessment Battery for Children-2 (KABC-2; Kaufman & Kaufman, 2004), Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1996), Stanford-Binet Intelligence Scale (Stanford-Binet IV; Thorndike & Hagen, 1986), Universal Nonverbal Intelligence Test (UNIT; Bracken & McCallum, 1998), and Wechsler Intelligence Scale for Children 4th Edition (WISC-IV; Wechsler, 2003). The CogAT and NNAT are group-administered tests whereas the rest are individually administered. Table 2.1 summarizes the differences in directions and examples across these tests.

Common features included (1) the brief and vague explanations of an analogical rule, (2) the reliance on two or three practice items, and (3) the use of extremely simple and uninformative practice items. The common misunderstanding of analogies in young children (Goswami, 1991) is not surprising given some of the demonstrations of analogies given in the published tests.

Part of the challenge of writing directions is how to describe clearly an abstract task in specific but general language—a problem compounded when examinees are quite young. How do we typically tell examinees how to solve an analogy? We ask them to “figure out what goes with C in the same way that A goes with B” or “choose the picture to finish the pattern. Try and work out how the pattern goes” (Goswami, 1991, p. 8). Even to an experienced test taker, these instructions are too ambiguous to know whether the task requires series completion or analogical reasoning. Examples are then crucial for showing rather than telling the examinee what the task is. However, many of the test directions surveyed in Table 2.1 appeared to assume that one or two simplified sample items would convey the task adequately. This seems overly optimistic. For a truly naïve test-taker, even two good examples is probably not sufficient to figure out what the task is when it is one of the unusual formats used in ability tests. Without a clear concept of

Table 2.1 Directions and examples from published tests

Test	Subtest	Explanation of the task (with indicated gestures)	Number of examples	Notes on examples	Feedback for one example
NNAT (group test)	Progressive matrices	“There is a piece missing in this puzzle. One of the answers goes on the question mark to finish the puzzle.”	2		“Number 4 is the answer because a piece that is all blue will finish the puzzle.”
WISC-IV	Matrix reasoning	“Look at these pictures. Which one here (<i>point to options</i>) goes here (<i>point to question mark</i>)?”	3	Does not show A to C transitions	“All of these butterflies are blue. This one (<i>point to answer</i>) is also blue, so it goes here.”
cTONI	Geometric Analogies	“(point to first picture) This is to (pt to 2 nd) this (pause) as this (point to 3 rd) is to which of these (run finger over pictures at bottom). Point to your answer”.	3	Shows A to B and A to C transitions in items	“If correct: head nod. If incorrect: examiner shakes head ‘no’ and moves examinee’s finger to correct answer.”
K-ABC	Conceptual thinking (odd-man-out for four pictures in matrix)	“Look at these pictures. One doesn’t go with the others. Point to the one that doesn’t belong.”	2		“These three are pigs and this one is a duck. The duck doesn’t belong.”
	Pattern reasoning (sequences)	“Look at this row of pictures. One picture is missing. Which one of these goes here?”	2		“The correct answer is this one. The solid green circle (<i>point</i>). The pattern goes green circle, green circle, green circle... (<i>point to each circle in turn</i>).”

Table 2.1 Continued

Stanford-Binet IV	Object Series/Matrices	<p>“Here is a big box with four little boxes inside. It’s like a puzzle that has one piece missing. The top two boxes show a pattern—the yellow triangle becomes a larger red triangle. The bottom two boxes must follow the same pattern, but one has a question mark on it.”</p>	1	<p>“This one should go in the box because it is a red triangle.”</p>
UNIT	Analogic Reasoning	No language, only gestures	4 sample items, 4 practice items	<p>If examinee catches on quickly, some may be omitted</p> <p>If correct: head nod. If incorrect: examiner shakes head “no” and moves examinee’s finger to correct answer.</p>
CogAT (Form 6; group test)	Figure Matrices	<p>“The big square is like a puzzle with one piece missing. You need to find the missing piece. Let’s do one together. What is missing? What goes in this empty square? ... Which of these pictures is the missing piece?”</p>	2	<p>“In this puzzle, all of the shapes are the <i>same</i>. The second picture is the same as the others in the puzzle.”</p>

the task, it seems unlikely that test takers would know which elements to attend to and which were irrelevant. From the related research, a few important features of examples became clear: (1) practice items must be unique to the format and not confusable with other tasks, (2) the examples must span the range of formats of the full test, and (3) the examples must be paired with verbal directions that draw attention to important elements. The type of feedback provided for examples is considered separately in the next section.

When only a few examples are provided, they should exemplify the format well and distinguish it from similar formats. Ross and Kennedy (1990) found that examinees used examples from instructional materials to generalize a procedure for later problem solving. Therefore, the range of examples provided is crucial for helping examinees develop appropriately generalized strategies (Klausmeier & Feldman, 1975). Test developers must also consider what other solution strategies test takers might develop from the examples, especially if the practice item could be an example of a more familiar task. For example, the *WISC-IV* had for its first practice item a matrix with three identical blue butterflies. The examinee was to select a fourth blue butterfly (not a red butterfly, green butterfly, or yellow butterfly) to complete the matrix. Based on this one example, what would a naïve test-taker think the task was? Perhaps they would assume it was a test of color matching, perhaps they would think it was a test of matching in general (i.e., you get three samples, and you have to fit in a fourth). Nothing about this example clearly indicates that the first two boxes are particularly linked, that the pictures are usually all different, or that there is a relationship to be deduced from the top row and applied in the second row.

Demonstrations of analogies given in individual intelligence tests often fail to provide a fully developed analogy in the examples—the only time where the examiner can give further assistance to examinees. Thus, it is not surprising that young children commonly misunderstand analogies (Goswami, 1991) and prefer distractors that simply match part of the stimuli (Vodegel-Matzen, van der Molen, & Dudink, 1994). In some

tests, there were no A-B transformations in the practice items, and others never show an A-C transformation. It is difficult to see how test-naïve students would come to understand that both A-B and A-C require some form of mapping if the examples never show such a transformation. In fact, one test went through all of the examples without ever showing the examinee that the A to B relationship was not always identity (i.e., it never showed the A and B terms as different pictures). Examinees may be able to figure out what to do when they first encounter an item without an A to C identity, but it is a possible source of bias if some examinees do anticipate such an item. To understand the task completely, examinees must be prepared to handle both transformations at once. Better test examples would use relationships that are clear and simple to any student (e.g., kitten:cat::puppy:?) but exemplify a true analogical relationship to the exclusion of any simpler relationship. Such examples might discourage the strategy of simply choosing an associate of the C term that is especially common among young students (Gentile, Tedesco-Stratton, Davis, Lund, & Agunanne, 1977; Goldman, Pellegrino, Parseghian, & Sallis, 1982).

The breadth of examples is also important. Examples can introduce strategies and allow examinees to anticipate the range of challenges in the test. Jacobs and Vandeventer (1971) showed that children were able to learn from examples and develop a basic solution strategy when a wide range of examples was presented. Performance on figure analogies (what they call double-classification items) improved when first-grade students received training on additional practice items compared to a control group taking only the pre- and post-tests. This training also led to improvement on a near-transfer post-test with different relationships to be discovered, but led to less improvement on a far-transfer task consisting of items from the Raven's Progressive Matrices.

Morrisett and Hovland (1959) also found that presenting a range of different practice items was essential because it teaches examinees to discriminate between task-relevant and task-irrelevant problem features. However, they also found limits to how

much variety is beneficial. Although they found that extensive practice on just one type of item led to negative transfer on diverse items, they also found that sufficient practice on a set of three item types was more effective than brief practice on 24 item types. Thus, the best approach in selecting practice items is to identify broad categories of items and help examinees develop workable but general strategies that can be adapted to different items.

Test developers should design directions and examples keeping in mind that students will rely more heavily on the examples to get a sense of the task. LeFevre and Dixon (1986) compared performance of students on figure series and classification tests when examples and instructions for test conflicted (e.g., the words described a series test but the example showed a classification problem). No matter how the experimenters varied the emphasis on the directions, students used the example as their primary source of information on how to do item. LeFevre and Dixon reported that their college-aged participants consistently disregarded the instructions and relied exclusively on the example.

However, LeFevre and Dixon (1986) warn that “the present conclusions probably would not apply to completely novel tasks; in such cases a single example might be insufficient to induce any procedure at all” (p. 28). They cite research that examinees need both examples to help concretize procedures for tests and instructions to help abstract the important features from the examples. Feist and Gentner (2007) concur, finding that verbal directions were essential to drawing attention to important features of test items, teaching basic strategies, and giving feedback on practice items. “In speaking, we are induced by the language we use to attend to certain aspects of the world while disregarding or de-emphasizing others” (p. 283). Examples *by themselves* may be less useful to naïve test-takers, but clearly, they play an important role in helping examinees develop specific strategies and procedures for solving test items (Anderson, Farrell, & Sauer, 1984).

Type of Feedback Provided

Practice or example items can play an important role in demonstrating the format rules to students. However, practice alone does not lead to optimal performance when it is not paired with appropriate feedback. The research indicates that (1) feedback makes practice more efficient for learning and (2) elaborative feedback that goes beyond yes/no verification improves learning, though too much feedback can actually impede learning.

Practice with feedback is more efficient. The importance of feedback was established early in the history of psychology (Thorndike, 1927). Morrisett and Hovland (1959) showed that feedback increased the amount of learning from each practice item, whereas providing no feedback resulted in the need for many practice items to produce the same amount of learning. Furthermore, practice without feedback has the tendency to make most students faster but not better at task (Thorndike, 1914, Ch. 14), because such practice leads to automatization of whatever strategy the student is using, whether efficient or not, and not necessarily improvement in performance. With feedback, practice does more because it gives the examinees information on the adequacy of their solution strategies.

Several studies supported the need to provide feedback for practice items. Tunteler, Pronk, and Resing (2008) conducted a microgenetic study of first-grade students solving figure analogies and showed that although all students improved their performance from practice alone, targeted feedback in a dynamic training session greatly increased performance. The positive effect was especially pronounced on items for which students could clearly state the correct analogical relationship in contrast to items on which they chose the correct answer but could not verbalize the rule.

Whitely and Dawis (1974) investigated the effect of various practice and instruction treatments on college-aged students' performance on verbal analogies. They used several treatment conditions including conditions in which students received practice on 50 items with and without feedback. Feedback was often (but not always)

related to higher scores compared to practice or instruction conditions. The effects were small, but the sample was college aged, so they may have been too old to benefit from instruction on how to solve analogy items.

Sullivan and Skanes (1971) compared the post-test performance of high- and low-ability students (aged 9 to 18 years) assigned to three conditions: practice plus post-test; pretest plus post-test; and pretest plus practice plus post-test. The tests consisted of letter series questions whereas the practice items consisted of *number* series questions. The practice consisted of first seeing a number series item with an explanation from the examiner as to what the rule was. That item was then followed by a letter series item that followed the same rule but the examinees were not given any instruction on them. The practice session included the full range of possible rules for the post-test. Sullivan and Skanes found that bright examinees did best on "pretest plus practice plus post-test" by a large margin, followed by "practice plus post-test." For low-ability examinees, "practice plus post-test" resulted in highest performance, though the differences between treatment groups were smaller. In this case, some explanations of strategies led to better performance whereas just seeing the items only led to gains for high ability examinees. This latter finding is consistent with research showing that high-ability examinees learn more from practice without feedback because they can more often correct their own errors (Shute, 2007).

The positive effect of feedback has been confirmed in cross-cultural research as well. Hessels and Hamers (1993) explored the use of a "train-within-test" paradigm for administering a battery of tests including a figure analogies format with a sample of 400 students in the Netherlands. Students were aged 5-6 and came from immigrant families from Turkey and Morocco. To increase the amount of training in the directions, Hessels and Hamers added additional practice items with basic correct/incorrect feedback. They also repeated items and provided demonstrations, though they did not offer specifics about their methods. Their methods reduced mean differences by 3-9 IQ points between

the Turkish and Moroccan samples compared to the Dutch sample (though mean differences remained 14 IQ points for both age groups). In a separate line of research in the Netherlands, Resing, Tunteler, de Jong, and Bosma (2009) also found that dynamic testing methods improved the performance of ethnic minority students completing a seriation task.

Ideal directions give feedback to the examinees on the reason for a correct answer on the sample problems. Several researchers have investigated what degree of feedback and explanations are optimal. In her review of the literature, Shute (2007) concluded that elaborative feedback provided in tutoring systems resulted in larger gains in comprehension compared to simple verification of the answer. In an earlier summary, Sullivan (1964) reviewed studies comparing focused instruction on items to undirected practice and found that direct instruction was preferable. “In more difficult transfer situations, that is, when the transfer problems are different from, and more complex than, the original problems, training in understanding general principles produces greater transfer gains than does simple memorizing” (Sullivan, 1964, p. 6). A classic example is the study conducted by Judd (1908) where he compared the performance of students throwing darts into water when one group received practice and another group received practice plus lessons on trajectory. The latter group performed better when the depth of the target was changed. This study demonstrated that teaching the fundamentals for building a strategy leads to better performance and that practice on a narrow range of tasks interferes with transfer.

A study by Kittell (1957) went further in comparing levels of guidance to provide in test practice in order to maximize the impact on later test performance. His study of sixth-grade students compared the amount of instruction on test items (of the verbal odd-man-out variety) and its effect on test performance. His most effective treatment was an intermediate amount of guidance where examinees worked through a series of problems given some guidance on the relationships to look for. By contrast, minimal instruction

gave no clues as to the type of relationship, and maximal instruction gave examinees both the relationship and the answers. The intermediate amount of instruction increased test performance on both a test with similar relationships in the items as well as a test with entirely different relationships. Giving examinees practice on specific strategies to solve the test made them not only better at solving similar problems but also, in Kittell's perspective, made them better at discovering new relationships. Unlike the Sullivan (1964) and Sullivan and Skanes (1971) studies, Kittell's highest level of instructions provided extensive structure and explanation and did not allow examinees to attempt items for themselves. Thus, feedback with explanations can be effective as long as the explanations are not too thorough and examinees have an opportunity to guess prior to hearing the feedback.

In my review of popular intelligence tests, elaborative feedback was often omitted. In all cases, the examinee was told whether the item was right or wrong, but only in about half of the tests was a specific justification offered. These justifications are likely essential in helping the test-naive examinee to quickly and accurately build an appropriately generalized schema for the task and its acceptable solutions. When the examiner says that "C" is the correct answer and nothing more, the examinee is left to infer what the reason must be. It cannot be assumed that it is always clear. Given how very simplistic these practice items tend to be, it seems unlikely that the examinee who did not understand the task before would be helped simply by knowing the answer.

Introducing Basic Strategies

Test directions with elaborative feedback can also introduce a basic strategy for examinees. Three basic strategies are particularly helpful for figure analogies: a constructive strategy, a systematic approach to solving double-rule items, and a verbalization strategy. All three strategies are useful to teach because they never give away an answer, but are helpful in solving items.

Malloy, Mitchell, and Gordon (1987) compared the effect of three treatments on the performance of college-aged students completing Raven's Progressive Matrices. The *cognitive training group* completed a pre- and post-test and received two hours of instruction with practice training them in useful strategies for solving the matrices. Their training activities introduced the constructive and verbalization strategies. The *exposure group* completed all of the same practice activities and tests as the cognitive training group, but without the training activities. Finally, the control group only completed the pre- and post-tests of parallel forms. They found that all three groups improved in performance from pre-test to post-test, but that the cognitive training group gained more (1.29 SD) than either of the other two treatment groups (0.80 and 0.23, respectively, effect sizes calculated by the author), suggesting that the constructive and verbalization strategies were useful in improving performance.

Budoff et al. (1974) introduced all three of these strategies in their study of a learning potential (or dynamic assessment) task with a sample of children aged 6-14. First, they had students draw in the answer for early items, leading them to attempt to *construct* the answer analytically before looking at the options. This strategy has been found to relate to higher performance on spatial tasks (Bethell-Fox, Lohman & Snow, 1984; Malloy et al., 1987). For young children, especially, encouraging children to construct an answer in their mind first may decrease their tendency to impulsively choose an early response option (Goldman et al., 1982).

Second, they helped students develop a strategy for solving double classification items, a variation that they found students did not often solve correctly. To teach these items, they had students follow and draw in each rule one at a time to construct the final answer in phases. This training led to better performance on double classification items in the posttest. "This procedure helped concretize the elements of the solution process so that many children, after this type of practice, could do the double classification problems mentally with very little trouble" (Budoff et al., 1974, p. 238).

Third, while showing practice items on overhead slides, the instructor called on students to indicate correct answer and reasons for their choice. This was likely effective because it led students to verbalize answers and model analytic solutions for each other. Verbalizing the rule for a reasoning task is an important and useful strategy for solving reasoning test items, especially analogies and even figure analogies. It is a strategy adopted by the most successful problem solvers (Goldman et al., 1982; Malloy et al., 1987). Language is known to have a strong influence on reasoning, especially for forming and holding in mind particular concepts (Feist & Gentner, 2007; Gentner, 1978, 2003; Loewenstein & Gentner, 2005). Internal speech particularly helps children avoid impulsive and spontaneous attempts at problem solving:

The specifically human capacity for language enables children to provide for auxiliary tools in the solution of difficult tasks, to overcome impulsive action, to plan a solution to a problem prior to its execution, and to master their own behavior. (Vygotsky, 1978, p. 28)

Thus, teaching students to put the analogy into words (and most helpfully to put a verbal label on the relationship between the elements of the analogy) may be particularly important to improving students problem solving.

Bridgeman and Buttram (1975) believed that it was this verbalization strategy that caused score differences between Black and White students taking nonverbal tasks like figure analogies (from Form 1 of the CogAT). They compared standard directions to an expanded version that gave more examples and encouraged a verbalization strategy with a sample of 4th and 5th grade students. Sample sizes were small, and results failed to reach significance (according to Humphreys' [1976] corrections to their analyses), but the pattern of scores indicated that mean differences were smaller in the treatment group. Humphreys went on to criticize any test that could be so easily modified by additional directions, but I think this underestimates the impact of format familiarity for young students taking any kind of test and how much a little extra practice can reinforce their use of appropriate strategies and their confidence in test-taking.

When teaching a verbalization strategy, an important consideration is that such a strategy may be more difficult to teach to students from low-SES or culturally diverse backgrounds. Students from these backgrounds are less likely to have this strategy of reasoning modeled at home and therefore are more dependent on in-school activities to develop these strategies (Hess & Shipman, 1965; Lubienski, 2000). These students may require more explanation and practice before adopting the strategy because verbalization, if not automatized by the examinee, creates an additional load on working memory.

Benefits of Teaching Strategies

The multiple demands on working memory during problem-solving (storage, retrieval, transformation) requires that one have a strategy for solving items. “This often means assembling a preliminary strategy on the basis of task directions and experience on the first few items, and then modifying this strategy as new items are encountered.” (Lohman, 2001, p. 6). For universally unfamiliar tests, developing solution strategies and inferring the rules of the task are part of the cognitive load of the test. However, this cognitive demand becomes a source of bias and extraneous to the test construct when it is not a part of the cognitive load for all students (Schnotz & Kurschner, 2007).

Assembling a strategy is more cognitively taxing than adapting a strategy. Assembly processes are activities where one must organize a sequence of actions to complete a task (Snow, 1992). These processes are important in novel or ill-structured tasks where new strategies and approaches must be generated (Lohman, 2000). Flexible strategy assembly requires a foundation strategy that the examinee applies to those items that are well within their span. From this perspective, the goal of test directions is to help examinees develop this basic strategy from which to build more refined strategies. Such a basic strategy should allow all examinees to be successful on the first easy items (usually referred to as teaching items). Once examinees have had this initial success, the more

capable will be able to develop strategies that are more sophisticated and adapt as the items increase in difficulty.

Universal Design Modifications Intended to Help ELL

Students

The goal of this study was to develop an improved set of directions that address important challenges when assessing ELL students and other culturally and linguistically diverse students. These directions would be used for all students with the goal of helping some students without diminishing the performance of students who are already well served by existing directions. Such enhanced directions should also be consistent with the Universal Design approach to test construction.

The concept of Universal Design began in architecture as a term describing public buildings that were originally designed to accommodate the physically handicapped rather than modified after construction to provide those accommodations. Such buildings had their accommodated and unaccommodated facilities seamlessly integrated so that everyone could easily access the building. In recent years, the premise of Universal Design has been expanded to the domain of instructional design as well as educational testing (Thompson et al., 2002). Test design from the Universal Design perspective attempts to develop tests in such a way from start to finish that it maximizes the ability of students to engage in and complete test tasks, reduces the need for accommodations, and assures that tests measure the intended constructs among diverse student populations (Ketterlin-Geller, 2008). Specifically, the Universal Design approach means that the tests are maximally inclusive of testing populations, show construct validity across groups, contain minimal bias, have clear and understandable test directions, are legible and visually appealing, and either facilitate or eliminate the need for test accommodations (Johnstone et al., 2008).

The benefits of Universal Design for testing are not limited to ELL students. Such Universal Design approaches often help low SES, ethnic minority, and cognitively impaired examinees to engage in the test. The goal is not to give an advantage to any group but to reduce bias and the need for accommodations in test administration.

Universal Design has been applied to item construction through simplifying language in the hope of increasing test performance of ELL students (Abedi et al., 2009). Universal Design in the form of linguistic simplification has also been discussed with reference to test directions, but less work has been done here comparing the effects of the modified and original directions. The assumption has been that fewer words are always better. However, research by linguists indicate that this may not be true (Yano, Long, & Ross, 1994). More research is needed on the effect of simplifying directions and its effectiveness for improving the performance of ELL students.

Simplified Language Modifications

Accommodations are another means of adapting the testing situation to the needs of ELL students. The most common types of accommodations made for the test directions are simplified directions, repeated directions, native language translations of directions, and reading written directions and test questions aloud (Butler & Stevens, 1997; Rivera, & Stansfield, 2003; Rivera, Stansfield, Scialdone, & Sharkey, 2000). If simplified directions have no negative impact on ELL students, test developers might simplify the directions for all students as part of a Universal Design modification. Ortiz and Ochoa (2005) state the problem this way: “Tests that are often seen as representing verbally reduced functioning may contain lengthy and possibly confusing verbal directions that can affect an individual’s ability to comprehend what is expected or to provide an appropriate response” (p. 160).

There is little research on whether simplified directions have a positive or negative effect on student test performance. However, there are several guides to creating

plain-language tests from CCSSO (Kopriva, 2000, 2008) and other groups (Gaster & Clark, 1995; Johnstone et al, 2008; Yano et al., 1994). Most recommend the following rules:

1. Keep sentence structures brief and straightforward, avoiding complex structures (though the linguists [Yano et al., 1994] say that this is not always optimal).
2. Use consistent and straightforward paragraph structure. “It is common for these students [ELLs] to be unable to recognize the item’s requirements (e.g., what the item is asking them to do) when the item is presented in more complex stems, sentences, or paragraphs” (Kopriva, 2000, p. 34).
3. Use present tense and active voice whenever possible.
4. Avoid rephrasing or rewording the same ideas, use the same name each time and avoid using pronouns. This includes avoiding ambiguous pronouns with unclear antecedents (e.g., they, it).
5. Use every day, high frequency words.
6. Avoid idioms or colloquialisms and words with multiple meanings.

Other modifications have stronger research support. First, test directions must minimize the number of basic concepts used when giving directions, especially to young students. In traditional test administration, basic concepts are essential to explaining the task, especially when a practice item is presented. Telling the examinees to “look at the *top row* of pictures” or asking, “What happens *next*?” both involve basic concepts that young students may not know. As a result, the directions contain many more basic concepts than one would expect or desire on a test for young students (Kaufman, 1978). Flanagan, Kaminer, Alfonso, Rader (1995) found that five major intelligence tests used unfamiliar basic concepts repeatedly in their directions. They concluded that comprehension is likely impeded by the use of difficult basic concepts, long sentences, and passive voice.

Second, orally administered test directions should reflect the best practices of how native speakers talk to nonnative speakers to increase comprehension. Yano et al. (1994) found that a native speaker addresses a non-native speaker, they speak more slowly, use more careful articulation, stress key words with pauses before and after, and use more full nouns and fewer contractions. Native speakers also use questions more often as a way of initiating topics, more repetition (including semantic repetition, or paraphrase), and more comprehension checks.

Nonverbal Test Directions: Pantomime and Gestures

The trend in abilities assessment seems to be towards using the shortest instructions possible. This may not be the best method for closing the test score gaps between ELL and non-ELL students. Shorter directions—where repetitions, explanations, and logical conjunctions are dropped—may actually increase the burden of English comprehension for examinees (Baker, Atwood, & Duffy, 1988; Davison & Kantor, 1982). Increasing comprehensibility is a complicated process. Using number of words as the sole criteria may cause more problems than it solves (Yano et al., 1994).

The admonition to simplify language is taken to its logical, though extreme, conclusion in calls for nonverbal directions (McCallum et al., 2001). There have been several efforts to develop alternative test directions that use pantomime to reduce or eliminate the need for language. However, even when pantomime or gestures are used to give directions, cultural loading can remain because nonverbal communication can be almost as culturally loaded as language (Oller, Kim, & Choe, 2001). In such a case, test validity depends on the examiner and examinee's ability to communicate nonverbally. Nonverbal tests are therefore "by no means the 'answer' to the issues being addressed" (Ortiz & Ochoa, 2005, p. 160).

The greatest concern with nonverbal directions is that language is essential in conveying precise directions to examinees. Theorists have argued that language plays a

crucial role in teaching and understanding new concepts (Gentner, 2003; Vygotsky, 1978). Nonverbal classroom instruction would be a ridiculous proposition for the average classroom. If the test directions are also a learning task, then it should be clear that language is an essential component of directions. There are many components to be explained, especially for novel tasks included on ability tests. Though perhaps some things could be conveyed with gestures and pictures, it is more efficient to teach many of these things with words.

One crucial function of language in instruction and test directions is reification (Gentner, 2003). Using a relational term can reify an entire pattern, so that new assertions can be stated about it. A named relational schema can then serve as an argument to a higher-order proposition. In test directions, it is often useful to name the task—for example, for number series, telling students they are looking for a “pattern” helps the examiner to describe clearly what patterns are acceptable solutions. Another function of language is to help examinees understand important dimensions of performance. For example, Detterman and Andrist (1990) found it difficult to help even college-aged student understand when a task was speeded using nonverbal directions.

Though little research has been done on the validity of nonverbal directions for ELL students, there has been extensive work done with deaf and hard-of-hearing children. Sullivan (1982) compared the WISC test performance of deaf and hard-of-hearing students when they were administered pantomimed instructions, pantomimed instructions plus visual clues (as suggested by Neuhaus, 1967, and others), and what she called “total communication” methods of directions. The pantomimed instructions “entailed the statement of the standard directions for each subtest and the execution of pointing movements as directed in the manual” (Sullivan, 1982, p. 781). The total communication treatments used the same directions and pointing as the pantomimed directions but also added the simultaneous use of sign language to explain the task. Sullivan found that test scores improved significantly when examinees were presented

with the total communication format because this format relied on familiar modes of communication. Neuhaus (1967) similarly found that examinees had difficulty making correct inferences about the test task when only pantomime directions were used. In Neuhaus's study, the use of sign language (a familiar language shared by examiner and examinee) led to improvements in scores of over 1 SD compared to the pantomime directions.

Sullivan (1982) argued that part of the problem with pantomime directions was unfamiliarity, even for deaf students. The pantomime method was not regularly used in the classroom, so it was not appropriate for testing. Anastasi (1937) likewise found from reports on the Army Beta that the use of pantomime and gestures were confusing to examinees because they were not the person's usual mode of communication. Examinees lost motivation in the tests because of the "artificiality of the situation produced by the elimination of language" (Anastasi, 1937, p. 491). Army examiners also found that it was difficult to standardize the directions when given in pantomime.

Nonverbal and language-reduced directions are limited in what they can show the examinee about the test. The tests best suited for nonverbal directions would be either simple or obvious tasks or familiar tasks that do not really require much instruction at all. When nonverbal directions are used, but the task is not familiar, the examinee must rely more on their ability to infer the rules of the task. This increases working memory load and the need to assemble a strategy without clues from the directions about what that strategy might be. For directions intended to provide basic strategies and specific directions for an unfamiliar test, pantomime directions are insufficient.

Limitations of Improving Directions

In attempting to improve directions, there are two concerns that should be considered. First is whether providing too much help in the directions will affect the

measurement of the desired test construct. The second is whether providing help will only benefit students who are already doing well (a Matthew effect).

Does Practice Increase or Decrease g-Loading?

A concern on some ability and aptitude tests is that the capacity to make sense of the task can actually be part of the intended construct. Thus, teaching strategies and guiding students to understand the task may actually eliminate construct-relevant variance for a test, particularly its general ability (*g*) loading. Research has demonstrated the crucial role that learning plays in defining the test construct. Carlstedt, Gustafsson, and Ullstadius (2000) found that when three different types of test formats were intermixed, the *g*-loading of items was diminished relative to a form where the three item formats were administered in groups. Their hypothesis was that this intermixing of item formats would increase the complexity and therefore the *g*-loading of the tests, but they discovered that, in fact, learning was a valuable contributor to measuring general ability (*g*). Lohman (2001) likewise found that “randomly ordering items on ability tests can make them poorer measures of *Gf*. ... In part, then, ability tests are themselves inductive-learning tasks” (p. 6). If the purpose is to assess general ability, then misunderstanding directions may be construct-relevant. However, Kvist and Gustafsson (2007) found that general ability was not as reliably measured when an examinee sample was culturally heterogeneous. They concluded that the relationship between tests and general ability was negatively affected by construct-irrelevant variance due to differences in opportunity to learn. In general, the research shows that a limited amount of familiarization increases the predictive validity of test scores (Biesheuvel, 1972; Dague, 1972; Maspons & Llabre, 1985; Sullivan, 1964). In her review of the literature, Ortner (1960) concluded that whenever validity was addressed in coaching studies on tests, the result was usually to increase predictive validity rather than to decrease it, as some have hypothesized.

Validity is maintained if coaching only allows examinees with relevant skills and knowledge to benefit from the test directions (Crocker, 2005).

Ultimately, so long as examinees are differentially familiar with the test, then there will be bias when comparing the scores of test-familiar and test-naïve examinees. For ELL students and other students who have disrupted prior education, it is reasonable to believe they have not had equal experience with tests. Thus, it is more ethical to attempt to mitigate this disadvantage, even at the expense of construct validity.

Matthew Effects

Another consideration in modifying the testing environment is whether the changes will only benefit students who already excel on the test. High ability students appear to gain more from test directions, especially when those directions are ambiguous or incomplete (Snow & Lohman, 1984). This effect is sometimes referred to as the Matthew Effect (Sullivan, 1964) because the “rich get richer” and gain more than low ability students. In other words, those high ability students who already do well on the ability test will do even better and gain more as a result of improving test directions compared to their low-ability counterparts.

Directions may act as a catalyst only for students who are *ready* to learn strategies. For strategies to be effective, examinees must recognize their value and be able to use them without overwhelming working memory (Schnotz & Kurschner, 2007). It may be that only students who come into the testing situation with some readiness for the demands of testing will be able to acquire the new strategies from a short training session. These examinees likely include bright students with little testing experience or students of average ability with familiarity with testing. The key is offering students strategies that are within students’ zones of proximal development (Vygotsky, 1978) and thus do not overwhelm working memory.

Several researchers have found evidence of a Matthew effect. Glutting and McDermott (1989) found that gains for high ability kids exceeded gains of low-ability when given test-wiseness instruction or skills training. Likewise, Sullivan and Skanes (1971) found that high ability students gained more from advanced practice compared to low ability students. However, Sullivan (1964) showed that the Matthew effect was more likely to occur with young, test-naïve examinees and when training focused on basic principles of the test problems (i.e., the focus of test directions). In contrast, the lower ability students gained more from training on basic test principles when the group of students was mostly familiar with the test. In such a group, bright students may have already developed these skills from previous test experiences and thus may benefit only from training on the test content (more like instruction than directions).

Kulik et al. (1984) conducted a meta-analysis of studies focusing on the effects of practice on aptitude and achievement test scores. They found that after one practice test, gains were larger for examinees of high ability than low, though gains were greater when the tests were identical rather than parallel. Kulik et al. also found that test-retest gains on identical test forms were greater for aptitude tests, but the same for aptitude and achievement tests on parallel forms. This latter finding, which was based only on a few studies, contradicts studies reported elsewhere which show larger practice effects on puzzle-like tasks such as the WISC (Ortar, 1972).

Conclusions About the Design of Test Directions to Improve the Validity of Tests with Novel Formats

Based on the literature review, I identified several substantiated guidelines for creating directions. It was clear that directions should be treated as an instructional activity that engages students in the task, use clear examples, provide feedback, coordinate visuals and description, and use age/level-appropriate language. Ideally,

directions will also be evaluated as a part of the validity argument for a test to assure that they adequately familiarize all students with the task.

Engaging children in the directions may require having students actively construct answers early on rather than passively listening to extensive directions prior to any practice. In fact, some research has shown that it may be better to let the students see examples before they are given any description of the task in order to make the description more concrete (Sullivan & Skanes, 1971). Other research has shown that training that encourages discovery of rules improves performance more than simply explaining or pointing out a large number of rules to students (Haslerud & Meyers, 1958; Kittell, 1957).

Directions should also provide relevant examples with feedback. Examples that are too simplified do not expose students to the full range of item types and may lead them to develop ineffective strategies for the task (Jacobs & Vandeventer, 1971; Morrisett & Hovland, 1959). Although the practice items should reflect the range of tasks presented on the test, it is not necessary to show many examples. Additional practice items provide diminishing returns.

Basic feedback alone has been found to improve performance over no feedback (Morrisett & Hovland, 1959). However, thorough feedback can lead to greater learning (Shute, 2007). Such feedback includes not only verifying that the examinee answered the sample items correctly, but also includes an explanation of *why* the answer is correct and why other options are not (Kittell, 1957). In addition, feedback can help shape the strategies students develop and give guidance on what a better strategy might be. In the case of figure analogies, some tests provide only practice items where the answer matches part of the item stem (i.e., a matching solution). Such simplistic examples do not provide students with enough exposure to the range of items they will encounter. Better directions would provide a range of relevant examples that require students to practice generating both single and double-rule solutions. Researchers have also shown that

presenting students with double-rule items greatly improves performance (Budoff et al., 1974).

Any efforts to simplify the language used should avoid unintentionally increasing the difficulty of understanding the task. In fact, more explanation may be necessary for students less familiar with the format in order to decrease the cognitive load of the directions. Likewise, any visuals that are included should contribute to concretizing and clarifying verbal directions. Visuals should enhance and not compete with the directions.

Finally, the most important lesson from the literature is that test directions are a vital contributor to test validity and should be evaluated for their effect on student test performance. Unless test developers and users are certain that all students have had equal opportunity to become familiar with the task, bias may be introduced through systematic differences in test familiarity.

Current Study

In the current study, I compare the effectiveness of three types of directions for explaining a figure analogies task to first- and second-grade students. One is the standard format, which provides two practice items and basic directions similar to those provided by published tests. The second (nonverbal-dynamic) is a reduced-language format that provides four examples with feedback, but no substantial increases in explanations over the standard format. The third treatment is a verbal-dynamic format that, in addition to using four examples, introduces several innovations that are intended to increase strategy acquisition and encourage more analytical problem solving. These innovations attempt to promote a constructive strategy, decrease impulsive responding, and discourage the development of the unproductive strategy where students choose a matching distractor instead of the option that satisfies a true analogical relationship.

Initially the study design called for a computer-based test in which the different types of directions could be randomly assigned to students. Computer presentation would

also permit combining Spanish and English in ways that better met the needs of each examinee. However, practical considerations required the use of paper and pencil tests with video-based directions shown on a class television. Such modifications have advantages and disadvantages. On the one hand, teachers are better able to manage student behavior when all students are completing the directions at once. On the other hand, it is unclear how well modern students learn from video presentations. Despite the prevalence of educational television programming, students are more familiar with receiving directions for interactive games and activities that take place in a computer environment. Thus, some unintentional novelty may be introduced by the use of videos to provide test directions. Other modifications made for the video-based test directions are explained in the following sections.

Multilingual Directions

Many states provide translated test directions or permit the use of sight translations (clarifications made on the fly by teachers) as accommodations for ELL students (Rivera & Collum, 2006). To provide translated directions to a portion of students in a classroom, some states permit the use of tape or CD players with headsets to provide translated directions. In this study, to increase the ease of administration and practicality for schools, the translated directions were interspersed with English directions as part of the video presenting the directions activities. This approach has been used previously with other CogAT studies and is consistent with a Universal Design approach to test development to the extent that it reduces the need for accommodations without helping students who do not need that assistance.

Combined Spanish and English directions can be important for the comprehension of many Spanish-speaking students gaining proficiency in English. Whereas their English proficiency may not be developed enough to support English-only directions, their Spanish proficiency in cognitive/academic language may also not be advanced enough to

fully comprehend directions (Cummins, 1980). Therefore, the best way to ensure that students fully understand the task demands is to give students an opportunity to hear the directions in both languages.

Off-loading Verbal Comprehension with Visuals

An important innovation in this study is the use of videos that provide dynamic visuals to support the oral test directions and paper-based testing materials. In typical test directions, the examinee must listen to orienting instructions (e.g., “look at the first example”, “look at the square”, etc.) as well as task-relevant instructions. The examinee must also actively integrate the incoming audio information (the spoken directions) with the visual information from the examples while ignoring irrelevant figures and examples. All of this processing adds to the student’s cognitive load. The goal of the video directions was to reduce extraneous cognitive load by highlighting pieces of the test booklet as necessary, hiding irrelevant pieces and examples until they are needed, and reducing the need for orienting directions. Thus, the directions can *show* rather than *tell* test-takers where to direct their attention. This approach reduces the amount of explanation required and the corresponding cognitive load.

Other researchers found improvements in learning when multimedia components were added to vocabulary-learning activities. Verhallen, Bus, and de Jong (2006) compared the effectiveness of computer-based reading materials with static versus moving images for a sample of language-learner students in the Netherlands. The audio for the book was identical across treatments. Their multimedia modifications to the book included simple changes such as zooming shots and sound to direct student attention to certain features. The researchers found that relatively minor multimedia additions led to significantly better comprehension for young students, presumably because the video drew students’ attention to important elements. In particular, students who saw the animated story recalled more implied elements (e.g., character emotions, goals

intentions), indicating that there was a particularly positive impact of multimedia on understanding abstract elements of the story. Such results support the argument that abstract elements of test directions will be particularly sensitive to the quality of visual cues provided and that multimedia can concretize these abstract elements better than static images.

CHAPTER III

METHOD

The purpose of this study was to investigate how the design of test directions affected the ability of examinees to engage in the test, particularly if they were unfamiliar with the task or had limited proficiency in English. Three types of test directions were assigned at random to the classrooms participating in the study. Directions were administered by video on a classroom television with both Spanish and English audio. Teachers guided the administration. The resulting design was a cluster-randomized experiment with an outcome variable (test performance) at the student level and a treatment variable (type of directions) assigned at the classroom level. I expected to find that scores on a nonverbal figure analogies task were improved by their assignment to two types of test directions that provided more extensive practice compared to a control set of directions.

Participants

The sample consisted of schools in a large suburban school district in the western U.S. that is home to a large population of Spanish-speaking ELL students. The total sample included 1,061 students in 46 classrooms. The average class size in the district was 23 with a range of 16 to 27 with the exception of one class, which had only 9 students. Schools had 3 to 5 classrooms per grade level. Eight schools participated, including three that tested both first- and second-grade classrooms. A breakdown of the sample is given in Table 3.1.

Public data released on the district website provided information about the diversity of schools and their performance on state achievement tests. Schools within the district varied in the size of their populations of ELL students. Some schools reported ELL populations as high as 50% in the first and second grades, whereas other schools reported fewer than 10%. The Hispanic population was quite large in all of the schools,

ranging from 24 to 80%. Performance on state proficiency tests varied as well. About half of the schools scored at or above the state average in percent proficient on the reading achievement test. On the math achievement test, all but one school scored above the state average in percent of students scoring at or above the proficient level.

The school district requested the participation of all schools with second-grade classrooms and offered voluntary participation for their first-grade classrooms. The first-grade sample consisted of schools with large ELL populations that volunteered in exchange for feedback on student performance. Younger students were selected because they are less likely to be familiar with the item formats and more likely to benefit from the test directions.

Table 3.1 Summary of sample

Treatment group	Number of classes		Number of students		Subtotal ELLs	
	Grade		Grade		Grade	
	1st	2nd	1st	2nd	1st	2nd
Standard	7	8	169	172	10	21
Nonverbal	8	8	183	198	42	38
Verbal	7	8	154	185	21	15
Total	22	24	506	555	73	74

In the participating schools, ELL students comprised 14% of the students. Across schools, the proportion varied from 6-51% ELL, with most under 15%. At the classroom level, three classrooms were composed entirely of ELL students, 10 classrooms had no ELL students, and the rest had an average of 13% ELL students.

Pilot Study Participants

The design of the directions and test were honed through pilot testing in a small school in Iowa with a large proportion of ELL students. At this school, six ELL and six non-ELL students completed the tests under different treatments. They watched the directions videos and completed items on laptop computers. I talked to the students after they completed the test to understand better how they solved items and to generate ideas for how to improve the directions. A few months later, I conducted additional pilot testing information on the revised video directions using a paper-and-pencil test with several non-ELL children.

Procedure

The types of directions used were (1) standard test directions, (2) nonverbal-dynamic directions, and (3) verbal-dynamic directions (described in detail in a later section). All three treatments were administered using video with English plus Spanish audio. This combination was chosen because the school district had a large population of Spanish-speaking ELL students, but many teachers were not able to read Spanish directions when needed. Although the directions were provided by the video, teachers led the testing activity and paused the videos frequently to give students time to think and to read parts of the directions scripts. Teacher scripts indicated places where the teachers needed to pause the video. In addition, the video contained audio beeps to signal pauses.

Teachers were trained on the study procedures by the testing coordinator for the school district. The coordinator explained the purpose of the study and showed the teachers a demonstration video to familiarize them with the procedure. The Directions for Administration provided more information about the study procedures with specific instructions for the directions format to which the teacher's class was assigned.

Test directions, led by the teacher with the video, took 5-10 minutes. Students then completed a 20-item figure analogy test. Teachers paced the students so they worked

through the test items at the same rate. This was intended to help students avoid rushing or dwelling too long on any one item. The study took approximately 30 minutes to complete, but there were no set time limits.

Materials

A 20-item figure analogies (FA) test was developed for the study. See Figure 3.1 for an example item. Items were selected from a pool of new items generated for Form 7 of the Cognitive Abilities Test (CogAT; Lohman, in press). These items had been subjected to preliminary item tryout prior to this study. Items were chosen to provide a range of difficulty appropriate for first- and second-grade students. Based on previous test forms created from these items, an internal consistency estimate of .87 was predicted for a 20-item test.

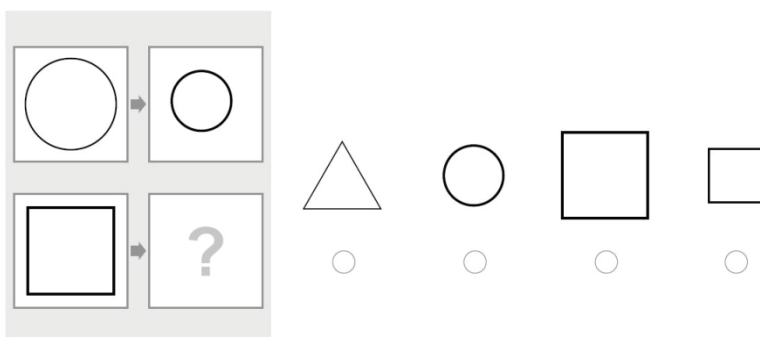


Figure 3.1. Example of a figure analogy

Based on the results of pilot testing, some items were modified to create more medium-difficulty items and to add options that would attract examinees who were using a matching strategy (i.e., attracted to any answer that looked like one of the figures in the item stem). The goal was to develop a test that would be sensitive to the experimental treatments but still show variability based on differences student ability. A variety of rules (tallied in Table 3.2) was used in the test items. These rules are representative of the

types and range of rules used on many figure analogies tests. Practice items included matching, size change, addition to figure, and double-rule (rotate and color change) items.

Table 3.2 Summary of item features

Feature	Number
Matching	3
Rotate/reflect	6
Split/combine figure	3
Addition to figure or size change	4
Double rule (rotate and feature change)	4

Previous research showed smaller practice or directions effects for more school-like tasks (Bergman, 1980; Evans & Pike, 1973; Glutting & McDermott, 1989).

Therefore, the figure analogies format was chosen because it was expected to be at least moderately unfamiliar for young students. Figural analogies and the related matrices format are widely used in ability testing (including the Raven's Progressive Matrices, UNIT, and NNAT-2) and are known to be strong measures of general mental ability. Thus, using this format for the test increased the generalizability of this study to published tests.

Additional Test Information

In the participating school district, second-grade students were administered the full Cognitive Abilities Test (CogAT, Form 6) less than one month after completing this study. I was able to match the study data to the CogAT scores for 457 out of 555 second-grade students who completed the study. The CogAT is a measure of fluid reasoning

abilities consisting of a Verbal, Quantitative, and Nonverbal Battery. At grade 2, each battery consists of 48 items in two formats. The Verbal and Quantitative Battery each consist of questions that are read aloud by the teacher with pictures for the answer choices. The Nonverbal Battery consists of two figural tasks (analogies and classification) with minimal verbal instructions. The Standard Age Scale scores (defined as $M = 100$, $SD = 16$) for the three batteries were used in the analyses.

Treatments

Three types of test directions were developed to explore the impact of directions on test performance. The standard directions treatment was designed to be representative of the instructions provided to students in published tests that use the figure analogies format. The verbal-dynamic directions were designed on the basis of pilot testing and previous research to provide the kind of discussion, practice, and strategy training that can be effective in encouraging young students to complete analogy items more systematically and analytically. Finally, the nonverbal-dynamic directions were designed to provide the practice given in the verbal-dynamic treatment but without the extensive discussion or training. Scripts for all three sets of directions are presented in Appendices A to C. Appendixes D and E present the student test booklets.

Standard Directions

To the extent possible, the standard directions were modeled on the directions used in existing tests (particularly the NNAT, WISC-IV, and Stanford-Binet IV) to promote generalizability of the results and assure that the treatment was a realistic representation of test directions actually in use. Students in the standard treatment saw two practice items with feedback. They then saw the two additional practice items used in the other treatments, but were not given any feedback on them, unlike the other two treatments. These “practice items” appeared as ordinary test items in this treatment..

Verbal-Dynamic Directions

The verbal-dynamic directions had several important features. First, they provided feedback and guidance on four practice items, compared to just two practice items in the standard directions. This provided practice on a more representative sample of test items. The two additional practice items exposed students to an additional single-rule item and a double-rule item so that strategies for both types of items could be modeled. Second, students were encouraged by the teacher to verbalize a rule for the puzzle (i.e., “the circle got smaller on the top row, so the square must get smaller in the bottom row”). This was intended to introduce a useful verbalization strategy as well as increase thoughtful over impulsive responding. Third, students were encouraged to draw their answers on the first two practice items and to imagine their answers for the last two practice items before looking at the answer choices for each practice item. This was intended to encourage a constructive strategy rather than a matching or option-elimination strategy (i.e., looking to the options for guidance). The constructive strategy is more commonly used by test-takers who are more able (Bethell-Fox et al., 1984).

Fourth, for the second and third practice items, the directions focused on each answer choice and explained why it was or was not the answer. Young students are particularly prone to selecting early options without looking at later options. This manipulation was intended to encourage thoughtful over impulsive responding by having students consider each option. For example, after having time to guess the correct answer for the second practice item, the students heard the following audio:

(Option a highlighted) Is this the answer? *¿Es esta la respuesta?*

No, the answer needs to be a square. *La respuesta tiene que ser un cuadro.*

(Option b highlighted) Is this it? *¿Y esta?*

No, it is smaller, but it is not a square. *Es más pequeño, pero no es un cuadro.*

(Option c highlighted) Is this the answer? *¿Es esta la respuesta?*

No, it is a square, but it is not smaller than the one in the puzzle.

Es un cuadro, pero no es más chico que el cuadro que está en el rompecabezas.

(Option d highlighted) Is this it? *¿Y esta?*

Yes, because it is a square that is smaller than the one in the puzzle.

Sí. Es un cuadro más pequeño que el cuadro que está en el rompecabezas.

This discussion was not used for the first practice item, because the pilot tests showed that students usually knew the answer immediately and would be bored by extensive explanations on such a simple item.

Fifth, for the final practice item, a double-rule example was presented along with a strategy for solving those kinds of problems. After the class engaged in discussion to name the two rules for the puzzle, students were told to circle the answer choices that fit the first rule, then to circle the choices that fit the second rule. Only the correct answer fit both rules. This is a strategy that could be applied to all of the double-rule items on the test and would help students keep both rules in mind while selecting an answer. Jacobs and Vandeventer (1971) found that keeping two rules in mind was difficult for young students to do without being taught a strategy to use.

The video for these directions included minimalist but key animations during the portions of the directions when students were looking at the video. Any time orienting directions were given (“look at the *top row*”), the relevant feature was highlighted on the screen. Any time item features were described in the video, a relevant animation was used. For example, in one case the rule is that the triangle had lines going *up and down*, so the video showed an arrow over the triangle moving up and down. A pencil marking the correct answer also appeared when the correct answer was revealed. This was intended to reduce the verbal loading of the directions to some degree.

Nonverbal-Dynamic Directions

The nonverbal-dynamic directions included all of the animations used in the verbal-dynamic directions and showed the same practice items with basic verification

feedback (correct/incorrect) so that students were given the same opportunities to see practice items and receive feedback on their answers. The directions differed importantly in that the vast majority of the descriptions of how to solve the items were omitted along with the teacher-led questions about the rules of the puzzles. No attempt was made to suggest a verbalization or constructive strategy.

Consistencies and Contrasts Across Treatments

Several aspects of test administration were held constant across treatments so that differences between groups would be due only to the directions format. First, all treatments were administered by the teacher in conjunction with a video played on a television. Second, all directions began with the same four sample items but with varying degrees of description and feedback. Third, all directions provided at least some animation highlighting the relevant parts of the practice item in the video corresponding to the audio. Fourth, all directions were provided in English and Spanish by the same reader. Finally, all students took the same test items following the directions.

The important distinction between the verbal-dynamic and the nonverbal-dynamic treatments was that the latter did not encourage students to follow any particular strategy or provide discussion of the answer choices on the last three examples. The important distinction between the two dynamic treatments and the standard treatment was that the latter showed the third and fourth practice item but treated them as regular items with no feedback or discussion.

Translations of Verbal Directions

Scripts for each set of test directions were written in English first and then translated to Spanish. The scripts were then recorded by a single speaker to create the English plus Spanish audio. A team of two translators created the Spanish scripts. One was a professional translator who grew up in Mexico and was an advanced graduate student in linguistics at the University of Iowa. He was primarily responsible for the

written translations and used a “combination of decentered/back-translation, to ensure the accuracy and fidelity of the original document into the target language” (A. Heras, personal communication, January 21, 2009). The other translator was an advanced graduate student in the Department of Spanish and Portuguese who grew up in northern Mexico with American parents and was thus fluent in speaking both English and Spanish. She reviewed and provided feedback on the first translator's work. She also recorded the English and Spanish audio used in the directions videos.

Both translators were briefed by the investigator on the purpose of the translations and the need for language appropriate to the context and age of the participants. Once the translated scripts were recorded by the second translator, they were spot-checked for audio quality by a third Spanish speaker.

Benefits of Using Video Administration

Using videos to administer the test directions served two purposes. One was to standardize the audio and the accent of the English and Spanish directions across classrooms. Many teachers in the participating district were not proficient in Spanish and could not read the translated directions themselves. A second purpose of the video was to offload some of the language comprehension demands through visual support. That is, when the audio said to look at the puzzle on the screen, the video highlighted what to look at. For the two dynamic video treatments, more animation was used to concretize the verbal directions.

Design and Hypotheses

The study design was a cluster-randomized experiment where students were nested in classrooms within schools and the experimental treatment was assigned at the classroom level. Multilevel modeling is essential when data consists of clusters of observations and independent variables that are best conceptualized at multiple levels. In such cases, the assumption of independent and random error required by many common

statistical methods (including common ANOVA designs) is violated by clustering of observations into groups that have smaller variances than the total sample.

In this study, a multilevel model was appropriate because the students were clustered in classrooms and schools and therefore student performance might be more similar to the performance of other students in their class or school than to students in the sample as a whole. Accurate interpretation of treatment effects required controlling for these similarities. In addition, although student characteristics such as ELL status are clearly best conceptualized as student-related variables, the study treatment was assigned at the classroom level. Ignoring this detail and analyzing the treatment as a student-related variable as part of a single-level multiple regression might bias the conclusions of the analysis. In this case, the error variance would be underestimated. Modeling a second level of classroom clusters was more accurate. Finally, because the schools varied in their achievement level and demographic composition, it was likely that classrooms within schools would be more similar to each other than to other classrooms in the district. Modeling these similarities required a third level of the model to be specified to control these effects.

The final design for this study was a three-level model with the dependent variable (test performance) and ELL status measured at the individual level, variables for treatment assignment and grade level at the second-order classroom level, and fixed effects for school differences at the third-order school level. All analyses were conducted using the HLM 6.0 software (Raudenbush, Bryk, Cheong, & Congdon, 2004).

Snijders and Bosker (1999) distinguished between multilevel models in which the clustering of units is the focal point of study and models in which the clustering is a source of nuisance covariance. In this study, the second level of the model (classes) represents variance of interest—the similarity of students in a classroom due to receiving the same directions treatment—but also nuisance variance due to potential similarity of students caused by the teachers. In contrast, the third level of the model was included

only to account for nuisance covariance caused by classes in a school being more similar to each other than to classes in the overall sample. Explaining variance at this level was not of particular interest.

Multilevel modeling introduces two important assumptions in addition to those common to regression models (Snijders & Bosker, 1999). The first assumption is that group random effects (τ_j and ψ_{0j}) at levels 2 and 3 are independent and are identically distributed across groups. These random effects reflect variation in intercepts and level-1 variable coefficients across classroom and school units. Tests for this assumption include inspection of coefficient variances and distributions of residuals. This assumption relies on correct specification of the model, particularly introducing random variables at higher levels of the model to account for variation in coefficients at lower levels. The second assumption is that the random variables for all three levels (σ , τ_j , ψ_{0j}) are normally distributed. This assumption is satisfied if the residuals for level-1 and 2 variables are normally distributed and have constant variance (Snijders & Bosker, 1999). A large number of level 2 and 3 units usually guarantees these assumptions are met.

Heck (2001) suggested beginning the analyses by fitting a single-level model to the data with all of the multilevel variables treated as if they were individual-level. Fitting this model gave an indication of which variables seemed to be related to the dependent variable and an indication of model adequacy (Heck, 2001). Heck also advised inspecting the intraclass correlations (ICCs), which capture the degree to which sample units are more similar to each other than the sample as a whole (the restriction in variability within units). Heck suggested a cutoff of .05 for a meaningful degree of ICC. If the ICC at the second (classroom) level was greater than .05, this indicated that measurement errors due to grouping affected the level-1 variables and a single-level analysis would yield inaccurate estimates. If the ICC at the school-level was smaller than .05, it suggested that the school level units did not contribute significant variance to the FA test scores and that

its effects could be discounted. In a three-level model, the intraclass correlations were calculated as follows:

$$\rho_{\text{Level2}} = \text{level 2 variance} / (\text{level-1 variance} + \text{level 2 variance} + \text{level 3 variance})$$

$$\rho_{\text{Level3}} = \text{level 3 variance} / (\text{level 1 variance} + \text{level 2 variance} + \text{level 3 variance})$$

(Snijders & Bosker, 1999)

In addition to ICCs, the design effects can provide a useful metric for determining whether multilevel modeling is warranted (Muthén, 1999). The equations for design effects in a three-level model are:

$$\text{D.E.} = 1 + (\text{average cluster size}-1) * \text{ICC}$$

and

$$N_{\text{effective}} = Nn / \text{D.E.}$$

(Snijders & Bosker, 1999)

Research Questions and Hypotheses

Preliminary Questions

1. Do class or school effects lead to strong intraclass correlations at L2 and L3?
 - a. Hypothesis: Expect ICCs to be greater than .05, warranting multilevel modeling.
2. Is there variability in the intercept or ELL coefficient across classes?
 - a. Hypothesis: Expect intercept and ELL coefficient to vary across classrooms.

Primary Questions

1. Do ELL and non-ELL students differ in average performance on the figure analogies test?
 - a. Hypothesis: Expect a main effect of ELL status with non-ELL students scoring higher on the figure analogies task.

2. Do verbal-dynamic or nonverbal-dynamic directions improve test scores compared to standard directions?
 - a. Hypothesis: Expect verbal-dynamic to result in highest mean performance, nonverbal-dynamic to result in second highest mean performance.
3. Is there an interaction of treatment with grade level such that younger students show stronger treatment effects?
 - a. Hypothesis: Expect treatment effects to be larger for grade 1.
4. Is there an interaction of treatment with ELL status?
 - a. Hypothesis: Expect verbal-dynamic and nonverbal-dynamic treatments to narrow differences between ELL and non-ELL students by raising performance of ELL students.
5. Did variations in treatment implementation affect the effect of treatments?
 - a. Hypothesis: Variations in treatment implementation will reduce the effectiveness of treatments. (See later section on “Additional Variables Capturing Variations in Treatment Implementation”.)

Follow-up Questions

1. Do the verbal-dynamic or nonverbal-dynamic treatments reduce the number of errors on the double-rule items?
 - a. Hypothesis: Expect DIF for double-rule items such that these items are relatively easier for students in the verbal-dynamic and nonverbal-dynamic treatments compared to standard treatment.
2. If directions do not have a large mean effect, do the directions have a narrow influence on students of middle or high ability?
 - a. Hypothesis: Expect verbal-dynamic treatment to spread out scores of students if high ability students benefit the most.

3. Do the verbal-dynamic or nonverbal-dynamic treatments reduce the popularity of matching distractors on non-matching items?
 - a. Hypothesis: Expect verbal-dynamic treatment to encourage more systematic problem solving and decrease the frequency with which students select matching distractors.

Multilevel Model

The basic theoretical model included only ELL status at the first level. At the second level, a variable for grade and dummy-coded variables for the verbal-dynamic and nonverbal-dynamic treatments were entered for the intercept (ρ_0 , β_{00}) meaning they were expected to predict variations in individual scores and class means. Interactions of grade with the treatments were also expected (β_{04} , β_{05}). For the ELL coefficient (ρ_1), which represents the effect of ELL status on individuals' FA scores, only the dummy-coded treatment variables were modeled as predictors of that coefficient, consistent with the hypothesis that the dynamic-directions treatments would reduce the negative effect of ELL status on test scores. Grade effects or interactions were not expected.

$$\text{Level-1 Model: } Y = \rho_0 + \rho_1*(ELL) + \sigma$$

$$\text{Level-2 Model: } \rho_0 = \beta_{00} + \beta_{01}*(GRADE) + \beta_{02}*(VTRMT) + \beta_{03}*(NVTRMT) \\ + \beta_{04}*(GRADE*VTRMT) + \beta_{05}*(GRADE*NVTRMT) + \tau_0$$

$$\rho_1 = \beta_{10} + \beta_{11}*(VTRMT) + \beta_{12}*(NVTRMT) + \tau_1$$

$$\text{Level-3 Model: } \beta_{00} = \gamma_{000} + \psi_{00}$$

$$\beta_{01} = \gamma_{010}$$

$$\beta_{02} = \gamma_{020}$$

$$\beta_{03} = \gamma_{030}$$

$$\beta_{04} = \gamma_{040}$$

$$\beta_{05} = \gamma_{050}$$

$$\beta_{10} = \gamma_{100} + \psi_{10}$$

$$\beta_{11} = \gamma_{110}$$

$$\beta_{12} = \gamma_{120}$$

The intercept at level 1 was modeled as a random effect at level 2 and 3 because the mean FA scores for students were expected to vary by classroom and school. The ELL coefficient was also initially modeled as a random effect because I expected the study treatment to influence the magnitude of the effect of ELL status on FA performance in different classrooms (i.e., that the size of the ELL coefficient would vary by classroom). Treatment and grade effects at level two were treated as fixed effects at level three because their magnitude was not expected to vary systematically across classrooms. In addition to the theoretical model, other variables related to class features (size of class, number of ELL students) and variations in treatment implementation were tested. This is explained in more depth in a later section. For clarification, see Table 3.3, which relates the primary research questions to the terms of the model.

Fit statistics

Throughout the model-building process, the fit of the model was monitored in several ways. First, the number of iterations required to estimate the model coefficients was monitored because the difficulty of fitting the model is related to the amount of data available to estimate the parameter (Snijders & Bosker, 1999). For example, the data provide less information about variance components than fixed effects, so introducing those parameters can greatly increase the number of iterations required to reach convergence and reduce the fit of the model to the data.

Second, a chi-squared change test was used to assess improvements in fit that resulted from adding additional parameters. HLM 6.0 calculates a deviance value for each model. Assuming the models are nested, the difference in deviance across models can be interpreted as a chi-square value with degrees of freedom equal to the number of parameters introduced.

Table 3.3 Overview of research questions and model

Parameter	Verbal description	Connection to primary Research Questions (RQ)
$\rho_0, \beta_{00}, \gamma_{000}$	Intercept for individuals, classes, and schools	N/A
$\rho_1, \beta_{10}, \gamma_{100}$	Main effect of ELL status	RQ1: Expect a negative main effect of ELL status indicating that non-ELL students score higher on the figure analogies task
β_{01}, γ_{010}	Main effect of grade on class intercept	Validity check, expect positive coefficient consistent with growth from grade 1 to 2 (where grade 1 = 0)
β_{02}, γ_{020}	Main effect of verbal-dynamic treatment on class intercept	RQ2: Expect positive effect of verbal-dynamic treatment
β_{03}, γ_{030}	Main effect of nonverbal-dynamic treatment on class intercept	RQ2: Expect positive effect of nonverbal-dynamic treatment
β_{04}, γ_{040}	Effect of interaction of grade and verbal-dynamic treatment on class intercept	RQ3: Expect negative coefficient indicating that treatment effects are larger for grade 1 (variable = 1 for 2nd grade students in verbal group)
β_{05}, γ_{050}	Effect of interaction of grade and nonverbal-dynamic treatment on class intercept	RQ3: Expect negative coefficient indicating that treatment effects are larger for grade 1
β_{11}, γ_{110}	Main effect of verbal-dynamic treatment on ELL status effects	RQ4: Expect negative coefficient indicating that verbal-dynamic treatment diminishes effect of ELL status
β_{12}, γ_{120}	Main effect of nonverbal-dynamic treatment on ELL status effects	RQ4: Expect negative coefficient indicating that nonverbal-dynamic treatment diminishes effect of ELL status
$\sigma, \tau_0, \psi_{10}$	Variance of intercept by individual, class, and school	Preliminary question: Expect that there is variance to explain in class intercepts
τ_1, ψ_{10}	Variance of ELL effect (coefficient) by class and school	Preliminary question: Expect that there is variance to explain in ELL coefficients

Finally, changes in R^2 were used to assess improvements in fit obtained by adding additional parameters. The interpretation of these values for a three-level model is complicated by the existence of multiple R^2 that could be calculated. R^2 also only reflects the improvement in fit from fixed effects. However, Snijders and Bosker (1999) suggested that calculating the level-1 estimate was still valuable and provided the following equation:

$$\text{Var} = \sigma^2 + \tau_0^2 + \psi_0^2$$

$$R^2 = 1 - (\text{Var}_{\text{full model}} / \text{Var}_{\text{empty model}})$$

Item Analyses

To explore the impact of test directions beyond their effects on overall score, specific items were targeted for further study to evaluate whether the directions affected students' accuracy or choice of distractors on those items. Differential item functioning (DIF) and Differential Bundle Functioning (DBF) were used to compare the verbal-dynamic and nonverbal-dynamic treatments to the standard treatment. Separate analyses were conducted with the dynamic treatments as focal groups. The SIBTEST program was used to test for DIF and DBF using a latent ability trait as the basis to make comparisons of item behavior within reference and focal groups (Clauser & Mazor, 1998). SIBTEST was appropriate because it permitted both DIF and DBF analyses and it has been shown to keep type I errors rates low when there is a relatively large N and small number of test items, as I had in this study (DeMars, 2009). DBF was of particular interest because I expected that the four items relying on a double-rule solution (#16-19) would lead to DIF amplification. That is, I expected that small amounts of DIF on individual items would cumulate to large amounts of DBF across that item bundle (Douglas, Roussos, & Stout, 1996).

Data Entry and Missing Data

Data entry included recording student responses from the test booklets as well as teacher surveys on which the teachers were asked to report any issues with the testing or suggestions for improvement. Teachers also entered student information on the front covers of the test booklets. State ID numbers were used instead of student names, and gender and ELL status of the students were collected. The state-defined ELL proficiency levels were used: Non-English Proficient (NEP), Limited English Proficient (LEP), and Fully English Proficient (FEP). Students who were never considered ELL students were classified as Non-ELL. A summary of variables collected are outlined in Table 3.4. The only missing data observed was in student item responses.

Student responses to test items were scored right/wrong and tallied to produce an overall score. Items that students neglected to answer or that contained multiple answers were scored as omits. Completion rate was good with 96% of students completing at least 19 out of 20 items. Only 5 students completed 12 or fewer items and were dropped from further analyses. In addition, the quality of the test items was evaluated so that poorly functioning items could be removed from the analysis. All of the items functioned well and were retained for the full analysis.

Dummy coding

Several variables were converted to dummy codes prior to analysis. ELL status for the students initially consisted of the four levels described above. These levels were converted into a dichotomous variable (0=non-ELL, 1=ELL) by assigning a value of 1 to all students classified as Non-English Proficient (NEP) or Limited-English Proficient (LEP). All other students, including those classified as Fully English Proficient (FEP) were assigned a value of 0 for this variable. The treatment assignment at the classroom level was also converted to two dummy variables (0/1) for assignment to the verbal-dynamic treatment and for assignment to the nonverbal-dynamic treatment.

Table 3.4 Variables collected

	Method of collection	Notes
Level 1 – Students		
Figure Analogies Test	Student item responses	Scored each item 0/1, summed item scores
ELL status	Teacher coded on student survey	Based on state proficiency classifications
Gender	Teacher coded on student survey	Not used in the analyses
Level 2 – Classes		
Treatment	Assigned at random by testing coordinator	Dummy coded (0/1) into two variables: (1) Assignment to Verbal Treatment, (2) Assignment to Nonverbal Treatment.
Number of Students	Teacher survey	When teachers did not respond to this survey question, this was coded simply as a count of student booklets
Number of ELLs	Count of student surveys	Count of student-level teacher codes
Compliance	Experimenter coded	Coded 0/1 by experimenter after inspection of student booklets
DVD failure	Teacher survey/ Experimenter coded	Coded 0/1 by experimenter based on the teacher comments about problems with testing
Grade	Teacher survey	Coded 0 for 1 st grade, 1 for 2 nd grade

Additional Variables Capturing Variations in Treatment

Implementation

As in any study involving human participants, particularly in a school setting, experimental control of the treatments was limited by uncontrollable and unexpected factors. First, the use of DVDs burned on a computer led to a greater number of video failures than expected. Although all DVDs were checked prior to shipping to the school district, a significant number of classrooms used DVD players that could not read the DVD format used. In most classrooms, the DVDs worked or VHS tapes were used, but

15 out of 46 teachers (one-third of the sample) reported that they were not able to play the videos.

For classrooms where the DVDs failed to play *and* when the teachers contacted the investigator or the testing coordinator about the issue, teachers were asked to use the provided scripts to read aloud the directions provided by the videos. These scripts captured all of the verbal information and practice for students in the absence of the DVD. Not using the videos resulted in the loss of small amounts of visual demonstration and the combined English and Spanish directions when teachers were not fluent in Spanish. Other classrooms with DVD failure did not contact the investigator or the coordinator, so in those classrooms there was variation in whether the teacher used the provided script. The vast majority did use the provided script and only two teachers made the decision to skip the directions entirely and proceed to the test items (see the “treatment compliance” discussion below). A variable was created for DVD/VHS usage to capture any variation in treatment effect caused by use of scripts instead of videos.

In addition to the technology failures, there were variations in how the teachers administered the directions. Given age of the students, teachers played a crucial role in motivating students, keeping them on task, and monitoring their comprehension of the directions. In their training by the testing coordinator, teachers were encouraged to stop or back up the DVD/VHS when students were confused and to endeavor to motivate students. In the verbal-dynamic treatment, encouraging students to pay attention to the directions was especially important because there were multiple steps to the practice activities including drawing answers, circling options, and engaging in teacher-led discussions. In pilot testing, it was often necessary for the examiner to encourage students to engage in these activities because students would not immediately act on the video’s instructions. From an inspection of student booklets, it was clear that not all students followed along with the directions carefully and completed the drawings that were intended to promote strategy learning.

As a result of the apparent variations in study administration, I decided to evaluate “treatment compliance” for classrooms using observations of student behavior on the practice items to control for part of the variation in study administration in the analyses. After examining all of the student booklets in a classroom, I scored the class overall (yes/no) on four dimensions of student behavior:

1. Did most students answer the first two practice items correctly?
2. Did most students answer the last two practice items correctly?
3. Did most students draw in answers for the first two practice items?
4. Did some students circle options on the fourth practice item?

These dimensions were based on tasks that the directions asked students to perform. I expected that classes in the standard treatment would be classified as yes only on the first dimension since they were not given feedback on the third and fourth practice item and were not asked to draw or circle anything. I expected classes in the nonverbal-dynamic treatment to be classified as yes only on the first two dimensions since they were given feedback on all four practice items, but were not asked to draw anything. Finally, I expected classes in the verbal-dynamic treatment to be classified as yes on all four dimensions. Classes not fitting this pattern of behaviors were flagged using a single *compliance* variable in the data.

Six classrooms were identified as *not compliant* meaning they did not appear to provide the test directions as intended (or, in two cases, at all) and therefore their results would not be representative of the treatments. For this reason, these six classrooms were omitted from all of the analyses. Descriptive statistics with and without these six non-compliant classrooms are presented in Chapter 4. Of these six classrooms, four did not have functioning DVDs. Two of the classrooms were first-grade classes assigned to the nonverbal-dynamic treatment. Four of the classrooms were second-grade classes including two assigned to verbal-dynamic, one assigned to standard, and one assigned to the nonverbal-dynamic treatment.

CHAPTER IV

RESULTS

Review and Results of Pilot Testing

The first round of pilot testing was carried out in a local school with a large proportion of ELL students. Twelve students were observed in pairs as they completed the figure analogies (FA) task in the various treatments. At the time, the test directions and items were presented on laptop computers. My first observation was that students showed frustration when they were required to listen to a lengthy description of the item before they had a chance to guess the answer. This was evident for the standard and verbal-dynamic treatments. Second, it appeared that many students completed the practice items without really understanding the task. Following the practice activities, many selected answers based on matching or other erroneous strategies. Third, many students were not systematic in examining all options and verbalizing a rule even in the verbal-dynamic treatment that encouraged this behavior. Students tended to select early options that did not fit an analogical solution well.

In response to the first observation, impatience of the students, all of the treatments were changed to allow students to try the first item, which was quite simple, without much preamble and then provide description afterward. In response to the second observation, two additional practice items were added to the verbal-dynamic and nonverbal-dynamic instructions to give students more feedback prior to beginning the task and to provide a wider range of examples. Finally, in response to the lack of systematic solution strategies, I added an option-by-option description to the verbal-dynamic directions that went through each answer choice and briefly described why it was or was not the answer. This was intended to encourage students to examine each option and compare it to the rule they constructed from the stem.

Finally, in further literature review following the pilot study, I found Budoff's (Budoff et al., 1974; Corman & Budoff, 1973) work with practice items on the Raven's Progressive Matrices and decided to add a practice item with two rules (the figure rotated and changed colors) so that I could model an effective strategy for that type of item. Budoff et al. (1974) found that Hispanic ELL students (ages 6-14) did not tend to get these items right without explicit practice on them. I also incorporated their constructive responding approach into the first two practice items (see Appendix B).

Prior to the pilot study, I had intended to use a computer-administered test to simplify the scoring of the test. Afterward, I decided to use videos and paper and pencil tests instead because the novelty of using computers introduced too much variance in examinee performance. Variations in computer skills, interacting with other students completing the task, and being distracted during the video were all greater issues than I had anticipated. In addition, I expected that it would be difficult for teachers to coordinate testing when only a few computers were available in each classroom. Group administration would simplify the study administration for the teachers and ensure that students focused more on the test task.

Second pilot sample. After extensive revisions, I piloted the new verbal-dynamic directions videos with two new non-ELL students. Having students guess the answer first (before any explanation) and draw the answer into the analogy matrix worked well in terms of slowing them down enough to treat the practice items as learning activities. Encouraging the students to guess and then give corrective feedback also seemed effective. Based on these students, I also added a component to the directions for the teachers to pause the directions video and ask the students a series of questions about what the answer must be. I based these questions on the conversations I had with the students in both pilot studies after they initially attempted the items. These conversations often led to students gaining insight into the answer for several analogy items. I believed incorporating this guidance into the directions would be beneficial to all students.

Sample

Specifics of the sample were reported in Chapter 3. In total, 46 classes in eight schools participated in the study, but only 40 were included in the analyses. Six were eliminated due to inconsistent implementation of the study procedure (see section on “Additional Variables Capturing Variations in Treatment Implementation” in Chapter 3). The remaining 40 classrooms yielded a sample of 882 students. Of the remaining classrooms, some were unable to play the provided DVD or VHS and used teacher-read directions instead. A breakdown of classrooms in the various treatments is presented in Table 4.1.

Table 4.1 Classroom treatment assignment and DVD usage

Treatment group	Used video		Did not use video	
	1 st	2 nd	1 st	2 nd
Standard	4	4	3	2
Nonverbal	6	6	0	3
Verbal	5	4	1	2
Total	15	14	4	7

Item Analyses

Descriptive statistics for items were obtained using the IteMan (2006) analysis package. Analysis of the items indicated that they had an acceptable range of item difficulty. The items had a range of p-values between .12 and .89 with a median of .38 for first grade. At second grade, the range was .20 to .90 with a median of .56. Fifteen out of 20 items were in the ideal range of p-values (.3-.8) for both grades. Pt-biserial estimates mostly indicated strong discrimination with a median of .50 and range of .23-.69 in first grade and a median of .49 and range of .31-.64 for second grade. None of the FA items was judged to have serious flaws, so all were scored for further analyses.

Descriptive statistics by grade and ELL status are provided in Table 4.2. Second-grade students scored .76 SD higher than first-grade students. Overall, non-ELL students scored .41 SD higher than ELL students. Within grades, first-grade non-ELL students scored .58 SD higher than ELL students; in second grade, non-ELL students scored .41 SD higher than ELL students. Table 4.2 also presents the descriptive statistics for the full sample of 46 classrooms, including the six classrooms with variations in treatment implementation. Omitting these classrooms had minimal effects on means and SDs.

Table 4.2 Descriptive Statistics

Grade	Compliant classes only				Original data		
	ELL	N	Mean	SD	N	Mean	SD
1	0	376	9.0	4.3	413	8.9	4.3
	1	57	6.7	3.5	68	6.9	3.5
	Total	433	8.7	4.3	481	8.6	4.3
2	0	380	12.2	4.2	461	12.0	4.2
	1	69	10.5	4.2	74	10.5	4.2
	Total	449	11.9	4.3	535	11.8	4.3
Total		882	10.4	4.6	1016	10.3	4.6

The descriptive statistics indicate that there was a sufficiently low floor for the tests as the mean was more than 2 SD above the lowest score in most cases (a standard specified by Bracken, 2007). An inspection of histograms for both grade levels confirmed that there was a normal distribution of scores, though the first-grade sample exhibited a slight positive skew and the second-grade sample exhibited a slight negative skew. See Figure 4.1. The internal consistency of the test was strong with a Cronbach's alpha of .85 at both grade levels and standard error of measurement (SEM) of 1.78.

Descriptive statistics within treatment groups were calculated to determine whether distributions were consistent and would support further analyses. As with the

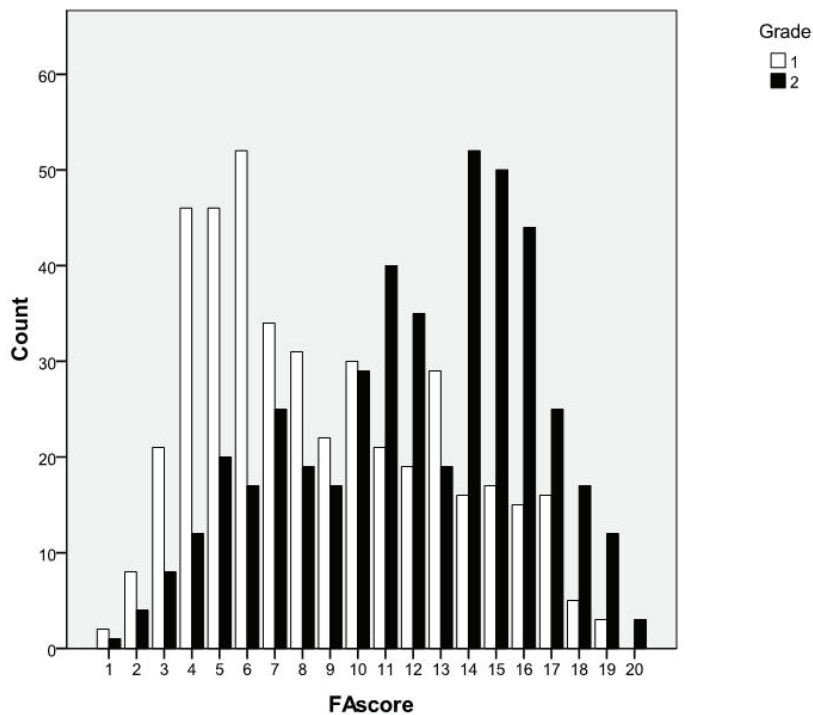


Figure 4.1 Distribution of scores on FA test

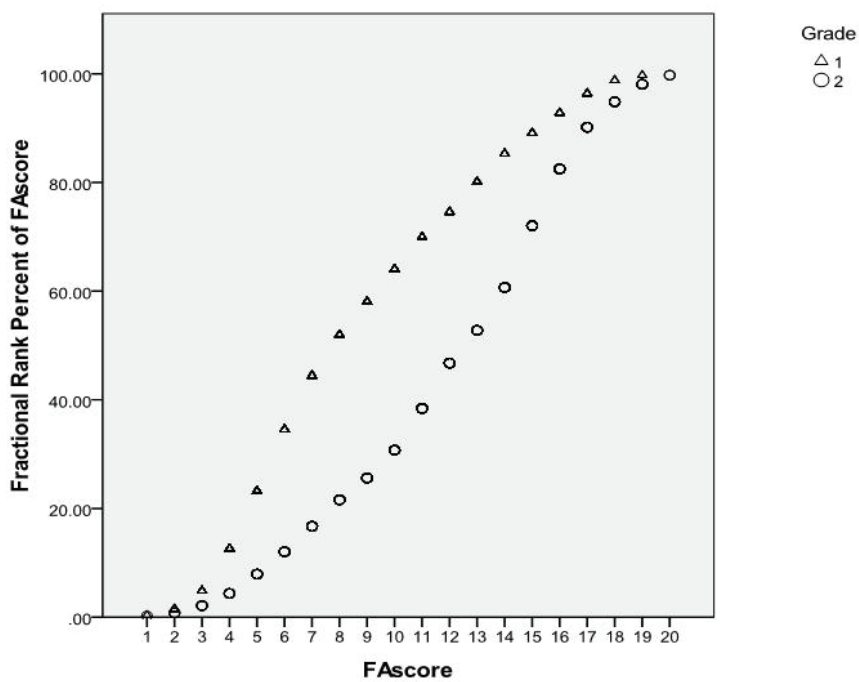


Figure 4.2 Plot of percentile rank by FA score indicates sufficient floor and ceiling

overall test (see Figure 4.2), the plot of percentile rank by FA score in Figure 4.3 showed little truncation of the scale at the either end of the distribution. The minimum range of 2 SD around the mean was not found for all treatments at grade 1 so there may have been a minor floor effect there. Overall, distributions within treatments appeared fairly normal without strong floor or ceiling effects.

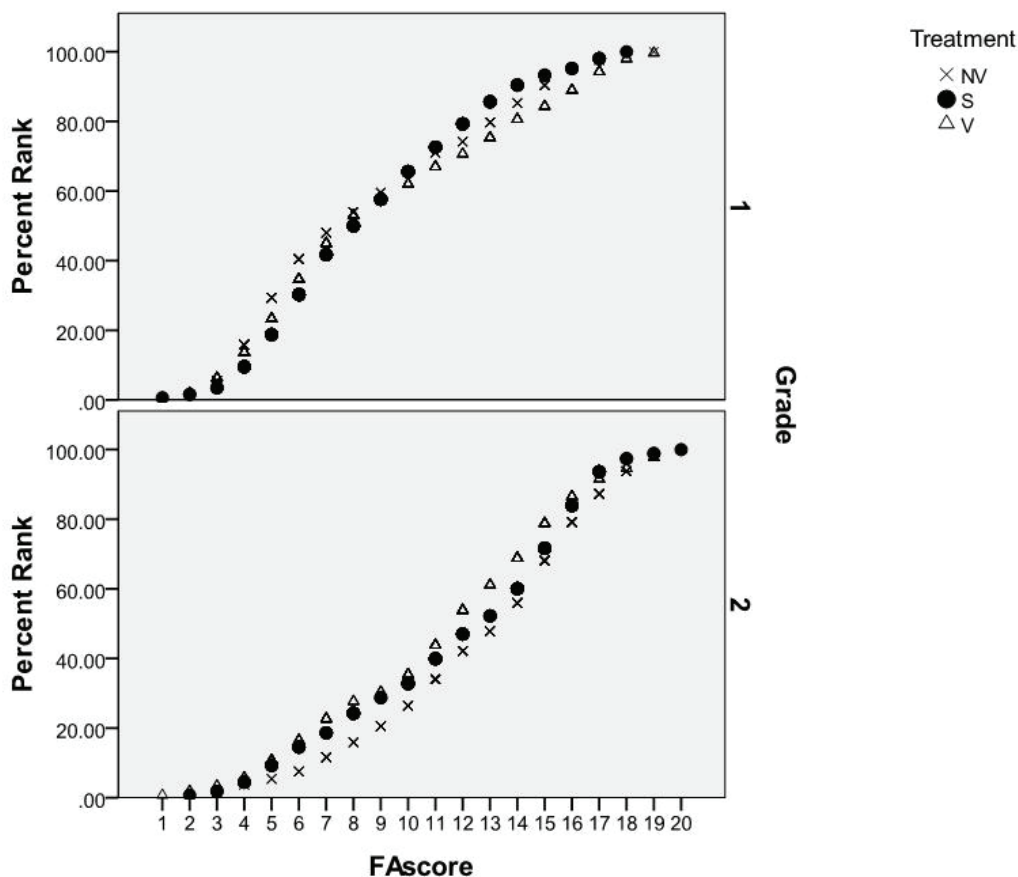


Figure 4.3 Percent Rank by FA score defined within grade and treatment

Preliminary Information About Treatments

Descriptive information about treatment effects within grades are presented in Table 4.3. Differences between treatments at grade 1 were all quite small with a .12 SD

effect between the highest (verbal-dynamic) and lowest (nonverbal-dynamic) treatments. Grade 2 showed the opposite pattern with nonverbal-dynamic having a mean .30 SD higher than verbal-dynamic.

Table 4.4 presents the means for the treatment groups across ELL groups. Between ELL groups, treatment differences were .42 SD for nonverbal-dynamic, .38 SD for standard, and .51 SD for verbal-dynamic treatment groups. Within the ELL group, the verbal-dynamic group scored lower than the standard group by .13 SD, whereas the nonverbal-dynamic group scored higher than the standard group by .09 SD.

Table 4.3 Test characteristics by treatment and grade

Grade	Treatment	N	Mean	SD
1	NV	126	8.4	4.4
	S	157	8.6	3.8
	V	150	9.0	4.7
2	NV	185	12.5	4.1
	S	134	11.8	4.3
	V	130	11.3	4.4

Table 4.4 Test characteristics by treatment and ELL status

ELL status	Treatment	N	Mean	SD
Non-ELL	NV	301	11.0	4.6
	S	286	10.3	4.3
	V	287	10.4	4.7
ELL	NV	79	9.1	4.5
	S	28	8.7	4.2
	V	35	8.2	3.9

Post-hoc Power Analysis

The software program Optimal Design (Spybrook, Raudenbush, Liu, Congdon, & Martínez, 2008) was used to estimate power under an HLM framework before and after collecting the data. Given actual sampling of this study (40 classrooms, ICC=14%), the power to detect effects was .80 for effect sizes of .38 and larger and .60 for effect sizes of .29 and larger. This indicates that, if the statistics for this sample are close estimates of the population values, the data were not sufficient to detect small differences in treatment effects. However, the data were sufficient to detect large mean differences such as those observed between ELL and non-ELL groups and between grade levels. The power to detect a difference in treatments at grade 2 of 0.30 SD was 0.60. For group differences in the population smaller than .30, type II errors may have occurred.

CogAT 6 Scores for Second-Grade Students

Three battery scores from the CogAT (Form 6) were available for 457 of the second-grade students in the sample. Compared to the national norms on the Standard Age Scale (M = 100, SD = 16), students in this sample scored below average on the Verbal Battery (M = 93.8, SD = 13.2) and Quantitative Battery (M = 94.9, SD = 14.2), but above average on the nonverbal Battery (M = 104.3, SD = 14.2). Correlations with the figure analogies test developed for this study were .43 with Verbal, .49 with Quantitative, and .65 with Nonverbal. Thus, the study instrument shows strong convergent and discriminant validity, especially with the CogAT6 Nonverbal Battery, which has both a figure analogies and a figure classification task.

Multilevel Model

To find an appropriate multilevel model, I followed a model-building process driven by theoretical and empirical questions and guided by the advice of Bryk and Raudenbush (1992), Heck and Thomas (2000), and Snijders and Bosker (1999). As a check on the validity of the proposed model, Heck and Thomas suggested that an HLM

analysis begin by fitting a single-level model to the data with all of the multilevel variables treated as if they were individual-level variables. Fitting this model gives an indication of which variables seem to be related to the dependent variable and an indication of model adequacy. I found that all of the following multilevel findings were consistent with the results obtained in a single-level regression. However, the incremental R^2 of adding treatment variables to the model was quite small in the single-level regression, indicating the treatments may not have a practically important effect.

The two crucial assumptions of multilevel models were also satisfied. The first assumption—that group random effects (τ_j and ψ_{0j}) are independent and identically distributed across groups—was satisfied because variance in level-1 coefficients within units was modeled with appropriate variables at level 2. The second assumption—that the random variables for all three levels (σ , τ_j , ψ_{0j}) are normally distributed—were satisfied because inspection of the level-1 residuals confirmed that they were normally distributed without problematically large tails. Snijders and Bosker (1999) specified these checks.

Step 1- Random coefficients model with ELL coefficient at level 1

$$\text{Level-1 Model: } Y = \rho_0 + \rho_1*(ELL) + \sigma$$

$$\text{Level-2 Model: } \rho_0 = \beta_{00} + \tau_0$$

$$\rho_1 = \beta_{10} + \tau_1$$

$$\text{Level-3 Model: } \beta_{00} = \gamma_{000} + \psi_{00}$$

$$\beta_{10} = \gamma_{100} + \psi_{10}$$

A random intercepts model was first fit to the data to estimate the intraclass correlations (ICCs) as an indicator of how similar students were within classrooms within schools (Snijders & Bosker, 1999). This model addressed the first preliminary research question: *Do class or school effects lead to strong intraclass correlations at L2 and L3?* The resulting level-2 ICC estimate was .142, meaning that 14.2% of the variance in individuals' FA test scores resided between classrooms. The ICC at level 3 was .052, so 5.2% of the variance in individual's FA scores resided between schools. In total, 80.6%

of variance resided between individuals and 19.4% resided between groups. At level 2, with 882 students in classrooms of 23 students, the design effect (DE) was 4.1 and the effective sample size ($N_{\text{effective}}$) was 214. At level 3, the design effect was 2.2 and the effective sample size was 410. Researchers suggest that an ICC of greater than .05 or a design effect of greater than 2 are grounds for pursuing a multilevel model (Heck & Thomas, 2000; Muthén, 1999). By both standards, there is evidence of relatedness between individuals within classrooms within schools that warrants modeling.

The second preliminary question addressed whether there was meaningful level-2 and level-3 variance in the level-1 parameters. Put differently, the question was whether estimates of intercepts and ELL coefficients varied greatly across classes or schools. Explaining such variance would be the focus of introducing explanatory variables at higher levels of the model. The random effects for the intercept at L2 and L3 (τ_0, ψ_{00}) were large and significant, indicating that there was variation in average test scores across classrooms and schools that could be explained using level-2 and level-3 variables. On the other hand, the random effects for the L1 ELL coefficient (τ_1, ψ_{10}) were nonsignificant. The lack of variability in the ELL coefficient (confirmed by the reliability estimate of .027 for that coefficient) indicated that classrooms and schools did not vary meaningfully in the size of the effect of ELL status on FA scores and that there was no variance to explain with higher-order variables. This result indicated that the fourth primary research question—*Is there an interaction of treatment with ELL status?*—was unlikely to be supported, as classroom variables did not appear to impact ELL effects.

Primary Questions

The first primary research question addressed was *Do ELL and non-ELL students differ in average performance on the figure analogies test?* The random coefficients model introduced above (“Step 1- Random coefficients model with ELL coefficient at level 1”) addressed this question by providing an estimate of the main effect of ELL

status on class means. The results (see Table 4.5) indicated that the effect of ELL status on FA scores was significant ($\rho_1 = \gamma_{100} = -1.90$, $SE = 0.41$, $CI_{95\%} -2.71$ to -1.09). This is consistent with the observation that there was a large (.41 SD) mean difference between the ELL and non-ELL samples.

Step 2 – Add grade at level-3 as a predictor of the intercept

$$\text{Level-1 Model: } Y = \rho_0 + \rho_1 * (\text{ELL}) + \sigma$$

$$\text{Level-2 Model: } \rho_0 = \beta_{00} + \beta_{01} (\text{GRADE}) + \tau_0$$

$$\rho_1 = \beta_{10}$$

$$\text{Level-3 Model: } \beta_{00} = \gamma_{000} + \psi_{00}$$

$$\beta_{01} = \gamma_{010}$$

$$\beta_{10} = \gamma_{100}$$

The next step in model building was to add a grade variable at the classroom level to control for the large .71 SD mean differences between grade 1 and 2. The model confirmed that there was a significant effect of grade on the intercept ($\beta_{01} = 3.25$, $SE = .53$, $CI_{95\%} 2.21$ to 4.30). This finding supports the validity of the tests, because the cognitive abilities measured by the test should grow rapidly at this age.

Step 3 – Add treatment variables at level 2 predicting intercept

$$\text{Level-1 Model: } Y = \rho_0 + \rho_1 * (\text{ELL}) + \sigma$$

$$\text{Level-2 Model: } \rho_0 = \beta_{00} + \beta_{01} (\text{GRADE}) + \beta_{03} * (\text{VTRMT}) + \beta_{04} * (\text{NVTRMT}) + \tau_0$$

$$\rho_1 = \beta_{10}$$

$$\text{Level-3 Model: } \beta_{00} = \gamma_{000} + \psi_{00}$$

$$\beta_{01} = \gamma_{010}$$

$$\beta_{03} = \gamma_{030}$$

$$\beta_{04} = \gamma_{040}$$

$$\beta_{10} = \gamma_{100}$$

Next, dummy variables were added to the model that represented assignment to the verbal-dynamic and nonverbal-dynamic treatments. This step of the model addressed the second primary research question: *Do different types of test directions improve test scores compared to standard directions?* Classes were coded 1 for VTRMT (β_3) if they received the verbal-dynamic treatment and 1 for NVTRMT (β_4) if they received the nonverbal-dynamic treatment. Classes in the standard treatment were coded 0 for both variables. The results indicated that the nonverbal-dynamic treatment had a positive effect on FA scores compared to the standard treatment ($\beta_4 = 0.65$, $SE = 0.36$, $CI_{95\%} = 0.05$ to 1.35), but the verbal-dynamic treatment did not appear to affect the classroom intercept compared to the standard treatment ($\beta_3 = -0.04$, $SE = 0.78$, $CI_{95\%} = -1.57$ to 1.49). Both dummy variables were retained in future models for consistency.

Although not shown in the tables, the fourth primary research question—*Is there an interaction of treatment with ELL status?*—was tested by adding a level-2 variable for treatment predicting the ELL effect at level 1 (ρ_1). As predicted, given the lack of variability in this variable discussed earlier, the effect was not significant and the variable was dropped.

Step 4 – Add grade- treatment interaction

$$\text{Level-1 Model: } Y = \rho_0 + \rho_1*(ELL) + \sigma$$

$$\begin{aligned} \text{Level-2 Model: } \rho_0 = & \beta_{00} + \beta_{01}(\text{GRADE}) + \beta_{03}*(\text{VTRMT}) + \\ & \beta_{04}*(\text{NVTRMT}) + \\ & \beta_{06}*(\text{GRADE} * \text{NVTRMT}) + \tau_0 \end{aligned}$$

$$\rho_1 = \beta_{10}$$

$$\text{Level-3 Model: } \beta_{00} = \gamma_{000} + \psi_{00}$$

$$\beta_{01} = \gamma_{010}$$

$$\beta_{03} = \gamma_{030}$$

$$\beta_{04} = \gamma_{040}$$

$$\beta_{06} = \gamma_{060}$$

$$\beta_{10} = \gamma_{100}$$

An interaction of grade and treatment was added to the model to address the third research question: *Is there an interaction of treatment with grade level such that younger students show stronger treatment effects?* I expected to find that younger students were more sensitive to differences in directions. Initially, when grade-treatment interactions were introduced for both verbal-dynamic and nonverbal-dynamic dummy variables (each variable had a value of 1 for second grade students in the dynamic treatment), neither was significant ($\beta_{05} = -0.86$, $SE = 0.96$, $CI_{95\%} -2.74$ to 1.02 ; $\beta_{06} = 0.83$, $SE = 0.63$, $CI_{90\%} -0.40$ to 2.06). However, since the verbal treatment was not significant to begin with, I also examined the grade-nonverbal interaction by itself, which was significant ($\beta_{06} = 1.23$, $SE = 0.45$, $CI_{95\%} 0.35$ to 2.12) and indicated that second-grade students gained more from being in the nonverbal treatment than first grade students. Since the latter model including only the grade-nonverbal interaction was empirically preferable and more parsimonious, I retained this model for additional tests. The grade-verbal interaction was dropped.

Step 5 – Add variables related to study administration as a predictor of intercept

$$\text{Level-1 Model: } Y = \rho_0 + \rho_1*(ELL) + \sigma$$

$$\text{Level-2 Model: } \rho_0 = \beta_{00} + \beta_{01}(\text{GRADE}) + \beta_{02}*(\text{DVD}) + \beta_{03}*(\text{VTRMT}) + \beta_{04}*(\text{NVTRMT}) + \beta_{06}*(\text{GRADE} * \text{NVTRMT}) + \tau_0$$

$$\rho_1 = \beta_{10}$$

$$\text{Level-3 Model: } \beta_{00} = \gamma_{000} + \psi_{00}$$

$$\beta_{01} = \gamma_{010}$$

$$\beta_{02} = \gamma_{020}$$

$$\beta_{03} = \gamma_{030}$$

$$\beta_{04} = \gamma_{040}$$

$$\beta_{06} = \gamma_{060}$$

$$\beta_{10} = \gamma_{100}$$

In step 5, I explored whether other available variables were related to class means. First, I entered classroom size and number of ELL students separately into the level-2 model; neither was significantly related to the intercept. Second, I entered a variable capturing whether the classroom used a DVD/VHS in administering the study. As detailed in Chapter 3, 15 classrooms were unable to play the videos provided and used teacher-read scripts instead. Thus, the implementation of the study procedure was altered and I believed this variation might reduce treatment effects or overall performance.

When DVD usage was entered as a direct effect on the level-2 intercept (the variable had a value of 1 when classes *did* use the video), it was a significant predictor. Surprisingly, the use of DVDs (i.e., correctly implementing the study procedure) was related to a *decrease* in test performance in the classroom ($\beta_{02} = -1.45$, $SE = 0.30$, $CI_{95\%} -2.03$ to -0.86). Explanations for this finding are explored in Chapter 5.

Step 6 – Add interaction of grade-DVD

$$\text{Level-1 Model: } Y = \rho_0 + \rho_1*(ELL) + \sigma$$

$$\begin{aligned} \text{Level-2 Model: } \rho_0 = & \beta_{00} + \beta_{01}(\text{GRADE}) + \beta_{02}*(\text{DVD}) + \beta_{03}*(\text{VTRMT}) + \\ & \beta_{04}*(\text{NVTRMT}) + \beta_{06}*(\text{GRADE} * \text{NVTRMT}) + \\ & \beta_{07}*(\text{GRADE} * \text{DVD}) + \tau_0 \end{aligned}$$

$$\rho_1 = \beta_{10}$$

$$\text{Level-3 Model: } \beta_{00} = \gamma_{000} + \psi_{00}$$

$$\beta_{01} = \gamma_{010}$$

$$\beta_{02} = \gamma_{020}$$

$$\beta_{03} = \gamma_{030}$$

$$\beta_{04} = \gamma_{040}$$

$$\beta_{06} = \gamma_{060}$$

$$\beta_{07} = \gamma_{070}$$

$$\beta_{10} = \gamma_{100}$$

For consistency with the other level-2 variables, I examined whether grade level interacted with DVD usage. This dummy variable had value of 1 for second grade classes that used the DVD/VHS video. The DVD-grade interaction was significant ($\beta_{07} = -1.25$, $SE = 0.76$, $CI_{95\%} -2.73$ to 0.23), but the main effect of DVD usage was reduced ($\beta_{02} = -0.71$, $SE = 0.56$, $CI_{95\%} -1.81$ to 0.39). This indicates that most of the effect of DVD usage on class means was due to a negative effect at second grade. The grade-DVD interaction was retained. An additional interaction of DVD use and treatment was introduced, but was not significant and therefore not retained in the final model.

The final model with only the theoretically and empirically supported effects was tested in step 6. See Table 4.5. Two of the effects were expected: students who were classified as ELLs scored about 2 points lower than non-ELL students and students in second grade scored 3.5 points higher than students in first grade. Other effects were more surprising. First, the verbal-dynamic treatment had no effect relative to the standard treatment. The only significant effect of treatments was a positive 1-point effect of nonverbal-dynamic directions for second-grade students. Also surprising was the negative effect of using the DVD to administer the directions. This effect was isolated to second-grade students and was associated with a 1-point decrement in performance. See Figure 4.4 for an overview of the effects in the final model.

Fit statistics for each step of the model-building process are presented in Table 4.6. Significant decrements in deviance (i.e., improvements in fit) were obtained by adding the ELL, grade, and DVD variables. The associated change in R^2 indicated that only adding grade level greatly improved fit. Other variables were associated with smaller improvements in variance accounted for. All of the models converged with a relatively small number of iterations, indicating that the data provided adequate information for estimating those coefficients.

Table 4.5 Coefficient estimates for each step of model building

Parameter	Random Coefficients						Full Model Step 6
	Empty model	Step 2	Step 3	Step 4	Step 5	Step 6	
Intercept ($\rho_0, \beta_{00}, \gamma_{000}$)	10.46 (0.51)	9.02 (0.34)	8.86 (0.59)	9.06 (0.63)	9.88 (0.53)	9.45 (0.31)	
ELL ($\rho_1, \beta_{10}, \gamma_{100}$)	9.46 to 11.46	8.36 to 9.69	7.71 to 10.02	7.83 to 10.29	8.84 to 10.93	8.84 to 10.06	
	-1.90 (0.41)	-1.93 (0.39)	-1.99 (0.38)	-2.01 (0.38)	-1.98 (0.34)	-1.98 (0.34)	
	-2.71 to -1.09	-2.68 to -1.17	-2.74 to -1.24	-2.76 to 1.25	-2.65 to -1.32	-2.65 to -1.32	
Grade (β_{01}, γ_{010})		3.25 (0.53)	3.19 (0.49)	2.74 (0.60)	2.78 (0.48)	3.58 (0.18)	
	2.21 to 4.30	2.24 to 4.15	2.24 to 4.15	1.56 to 3.93	1.84 to 3.71	3.23 to 3.94	
V trmt (β_{03}, γ_{030})		-0.04 (0.78)	-0.04 (0.78)	-0.03 (0.80)	0.06 (0.73)	0.00 (0.81)	
	-1.57 to 1.49	-1.57 to 1.49	-1.61 to 1.55	-1.37 to 1.49	-1.37 to 1.49	-1.59 to 1.59	
NV trmt (β_{04}, γ_{040})		0.65 (0.36)	0.65 (0.36)	-0.02 (0.53)	0.43 (0.49)	0.13 (0.65)	
	-0.05 to 1.35	-0.05 to 1.35	-1.06 to 1.03	-0.53 to 1.40	-0.53 to 1.40	-1.15 to 1.41	
DVD (β_{02}, γ_{020})					-1.45 (0.30)	-0.71 (0.56)	
					-2.03 to -0.86	-1.81 to 0.39	
Grade * NV (β_{06}, γ_{060})				1.23 (0.45)	0.78 (0.45)	1.04 (0.58)	
				0.35 to 2.12	-0.10 to 1.67	-0.11 to 2.18	
Grade * DVD (β_{07}, γ_{070})						-1.25 (0.76)	
						-2.73 to 0.23	

Table 4.6 Fit statistics

Variable added	Empty model	Random Coefficients				Full Model Step 6
		Step 2	Step 3	Step 4	Step 5	
		Grade	Treatment	Grd x Trmt	DVD	Grd x DVD
Variance estimates						
Level 1 variance	16.81	16.52	16.52	16.52	16.52	16.52
Intercept (τ_0)	2.95	2.91	0.91	0.86	0.69	0.62
Intercept (ψ_{00})	1.26	1.07	0.29	0.25	0.00	0.00
R ² change from empty model		0.02	0.15	0.16	0.18	0.18
Iterations	7	8	8	10	26	26
Deviance (df)	5062.91 (4)	5046.93 (5)	5013.42 (8)	5011.47 (9)	5002.71 (10)	5000.63 (11)
ΔX^2 Deviance (df)		-16.0 (1)	-2.1 (2) ^{NS}	-1.9 (1) ^{NS}	-8.8 (1)	-2.1 (1) ^{NS}

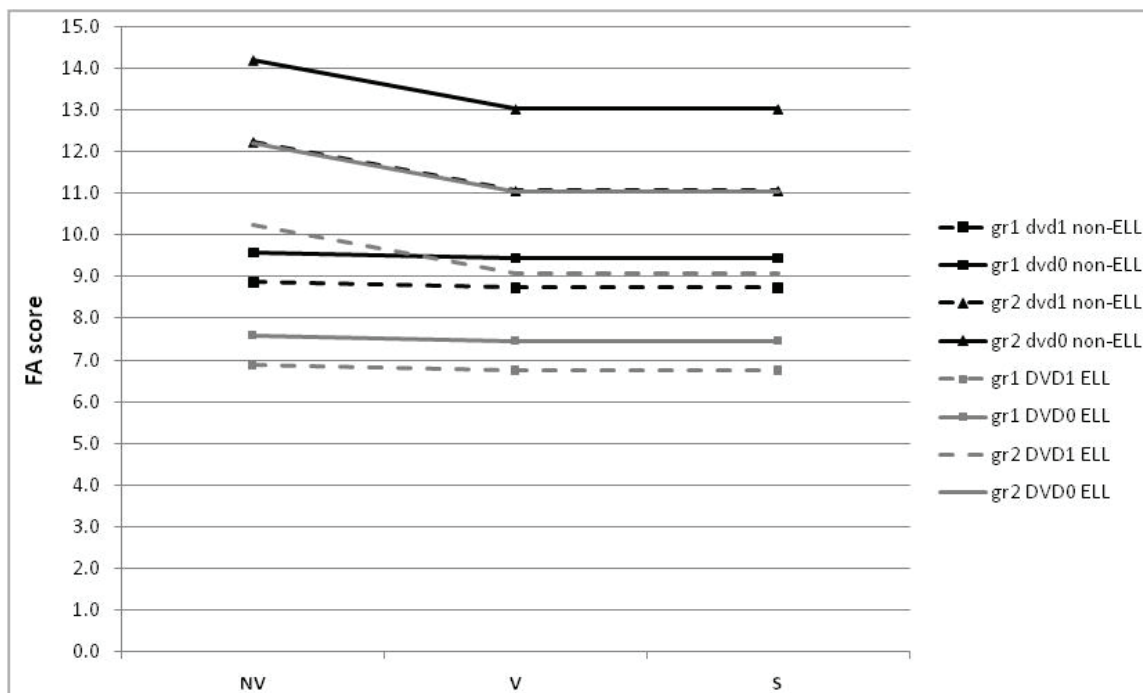


Figure 4.4 Overall Effects in Full Model. DVD0 = no DVD used

Differential Item and Bundle Functioning

Even though the verbal-dynamic treatment failed to have a main effect, I was interested in whether the dynamic treatments had an effect on particular items. In my first follow-up question, I asked—*Do the verbal-dynamic or nonverbal-dynamic treatments reduce the number of errors made on the double-rule items?* To address this question, I used differential item functioning (DIF) and differential bundle functioning (DBF) analyses to determine if the double-rule items (#16-19) were easier for either of these treatment groups compared to the standard treatment group. Table 4.7 shows the DIF and DBF results. To increase the sample size, the two grades were combined. For the verbal-standard comparison, small DIF was found on seven items including three of the double-rule items, though two of these effects favor the standard group. For the nonverbal-standard comparison, only two items showed DIF including one double-rule item showing medium DIF favoring the nonverbal group.

Table 4.7 DIF and DBF results

Items	Item rule	Beta est.	Std error	p-value	Beta est.	Std error	p-value
1	Matching	0.00	0.02	0.85	0.03	0.02	0.21
2	Matching	-0.02	0.03	0.55	-0.04	0.03	0.14
3	Matching	0.01	0.03	0.68	-0.01	0.03	0.77
4	Addition/size	0.03	0.04	0.50	-0.01	0.03	0.82
5	Split/combine	0.01	0.04	0.79	-0.01	0.04	0.71
6	Rotate/reflect	0.04	0.03	0.15	0.01	0.03	0.73
7	Rotate/reflect	0.01	0.04	0.84	0.06	0.04	0.16
8	Split/combine	0.00	0.04	0.95	0.00	0.04	0.94
9	Addition/size	-0.03	0.04	0.43	-0.02	0.04	0.64
10	Addition/size	-0.02	0.04	0.58	0.03	0.04	0.49
11	Addition/size	-0.05	0.04	0.21	-0.02	0.04	0.56
12	Rotate/reflect	0.03	0.04	0.51	0.04	0.04	0.34
13	Rotate/reflect	-0.03	0.04	0.46	-0.04	0.04	0.38
14	Rotate/reflect	0.07	0.04	0.06	0.03	0.04	0.43
15	Split/combine	-0.06	0.04	0.12	0.00	0.03	0.90
16	Double rule	0.05	0.04	0.28	0.02	0.04	0.66
17	Double rule	0.04	0.03	0.20	-0.01	0.03	0.72
18	Double rule	-0.06	0.04	0.15	-0.11	0.04	0.01
19	Double rule	0.01	0.04	0.71	-0.03	0.04	0.46
20	Rotate/reflect	0.01	0.04	0.85	-0.02	0.04	0.66
Bundles							
By item order							
1-3		0.00	0.07	0.98	-0.01	0.07	0.94
4-7		0.08	0.10	0.39	0.05	0.09	0.60
8-11		-0.13	0.09	0.15	0.00	0.09	0.97
12-15		0.01	0.08	0.92	0.05	0.08	0.56
16-19		0.05	0.07	0.45	-0.12	0.08	0.14
By rule							
1, 2, 3	Matching	0.00	0.07	0.98	-0.01	0.07	0.94
4, 9-11	Addition/size	-0.13	0.08	0.12	-0.05	0.08	0.58
5, 8, 15	Split/combine	-0.03	0.07	0.67	0.01	0.07	0.87
6, 7, 12-14, 20	Rotate/reflect	0.11	0.12	0.36	0.08	0.11	0.46
16-19	Double rule	0.05	0.07	0.45	-0.12	0.08	0.14
Matching distractors							
4-7, 12-14	Match for C term	0.23	0.14	0.11	0.08	0.13	0.54

Note: Positive beta means reference (Standard) group favored; negative beta means focal group favored; absolute values over .05 in bold

Item bundles were also considered because I suspected that items 16 to 19 might show an amplification of DIF as a bundle. Three other bundles of items were created based on item order for comparison purposes. The bundles mostly showed small and inconsistent effects, except for the focal bundle for the nonverbal/standard comparison, which showed substantial DBF. This indicated that these items were substantially easier for the students in the nonverbal-dynamic treatment. This is likely in large part due to DIF on item 18, but there is some amplification effect with items 17 and 19.

Items 8 to 11 also showed larger DBF for the verbal-standard contrast. The reason for this finding is unclear, especially given that DBF was also observed when many of the same items (9, 10, and 11) were classified by rule as addition/size change items. One explanation is that verbal-dynamic treatment students benefitted from the practice with feedback they were given on an addition-rule item. It is also worth noting that these items are the first to depart significantly in complexity from the basic rules presented in the early items. The DIF amplification of these items was pronounced as none of these items showed large DIF on their own.

Effects of Treatments by Quartiles

In my second follow-up research question, I was interested in whether treatments showed a Matthew effect—*If directions do not have a large mean effect, do the directions have narrow influence on students of middle or high ability?* One way of looking at this issue was to compare the average FA score of students within quartiles defined within grades and within treatments. Figure 4.5 shows a plot of these means. To illustrate, the top trend line in the graph shows the four means corresponding to quartiles of second-grade students assigned to the nonverbal-dynamic treatment (Q1 = 6.7, Q2 = 11.4, Q3 = 14.5, and Q4 = 17.1). Because classrooms were randomly assigned to treatments, students within the treatment groups can be considered randomly equivalent. Thus, differences in mean scores between students in the same quartile but different

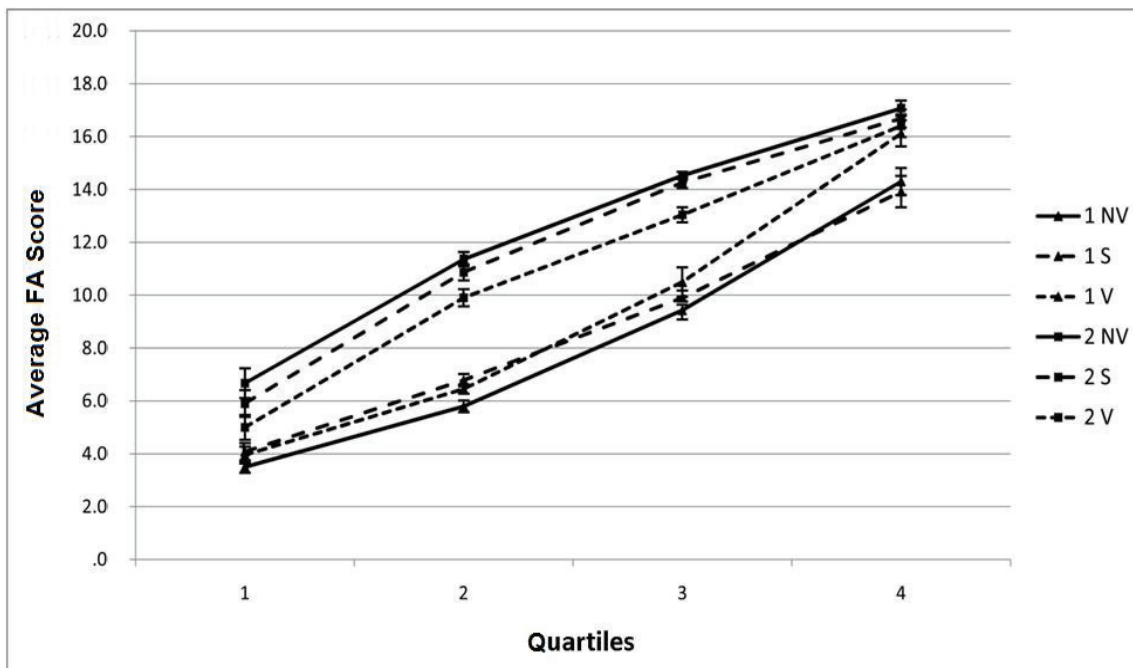


Figure 4.5 Mean FA scores with standard error bars for quartiles defined within treatment and grade

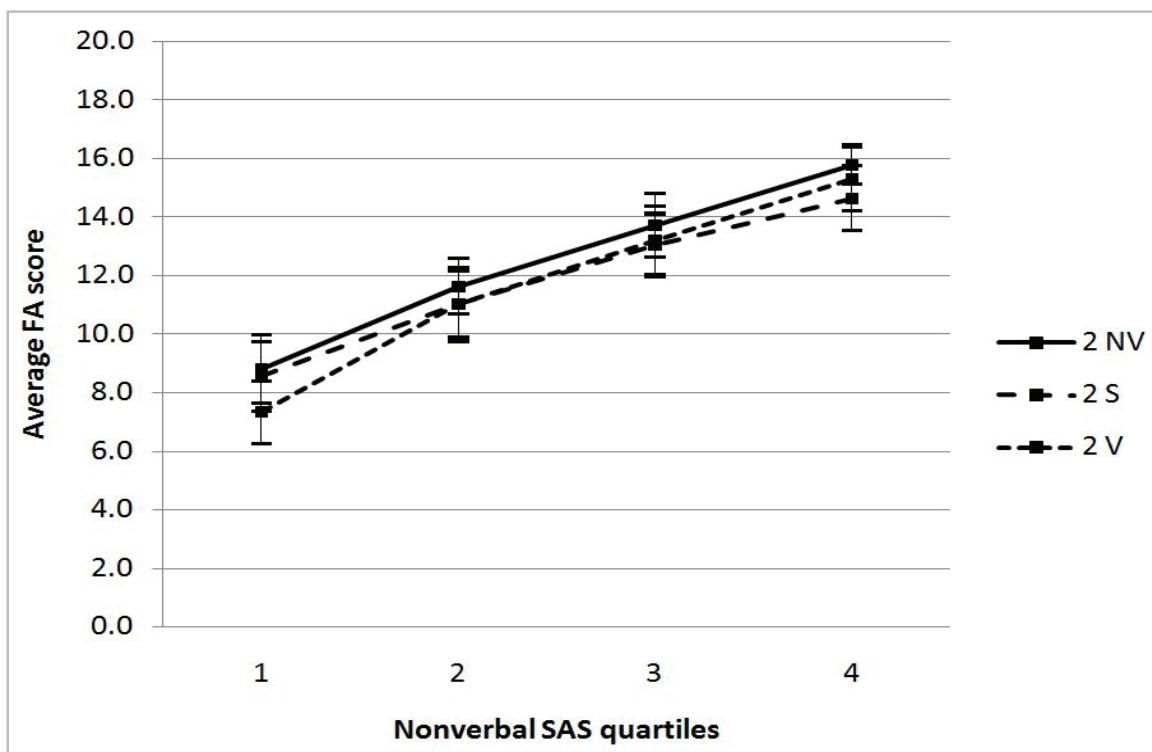


Figure 4.6 Mean FA scores for quartiles defined with CogAT 6 Nonverbal SAS scores

treatments can be taken as treatment effects within ability groups. The roughly parallel lines found for the nonverbal-dynamic and standard second-grade treatment groups is consistent with a mean effect for the nonverbal-dynamic treatment that shifts the four quartile means upward but is not consistent with a Matthew effect. On the other hand, the first-grade verbal-dynamic group has a noticeably steeper slope of means compared to the other first-grade groups. This indicates that the four quartiles in that group are spread out more, possibly indicating a greater advantage of the treatment for high-ability students. It also indicates that the most able students in that group scored higher than the most able students in the other first-grade treatment groups. In fact, the most able first-grade students in the verbal-dynamic treatment scored as well as the most able second-grade students. The 95% confidence bands based on the standard error of the mean in Figure 4.5 confirm this finding. This may indicate that higher ability students gained more from this treatment than the students in the other quartiles of ability. In second grade, there is a smaller, though similar trend, where the verbal-dynamic treatment group scored significantly lower in the 1st to 3rd quartiles, but equally well in the 4th quartile compared to the other second-grade groups.

Another point of comparison was the CogAT 6 scores available for part of the second-grade sample (457 students). After forming quartiles based on the CogAT 6 Nonverbal Battery Standard Age Scale, I plotted the mean FA scores with 95% confidence bands based on the standard error of the mean. The results in Figure 4.6 are consistent with those in Figure 4.5 for grade 2, though in this case the treatments have overlapping confidence intervals indicating they are not significantly different. However, the verbal treatment group appears to have a slightly steeper curve than the other groups at the lower quartile, indicating that the verbal treatment might have been problematic for lower ability students.

Variability in Treatment Effect

The preceding analyses address only whether treatments work well for the majority of students. I was also interested in whether there was evidence that a subset of students found the different directions particularly helpful or unhelpful. One way of looking at this issue was to find students who scored much better or worse than students from the same CogAT 6 quartile group and treatment group. Figure 4.7 shows boxplots of FA scores across ability quartiles based on the CogAT Nonverbal Battery scores. These plots identified eight students in the 3rd and 4th quartiles of CogAT Nonverbal scores who scored much lower on the experimental figure analogies test than other students in that quartile.

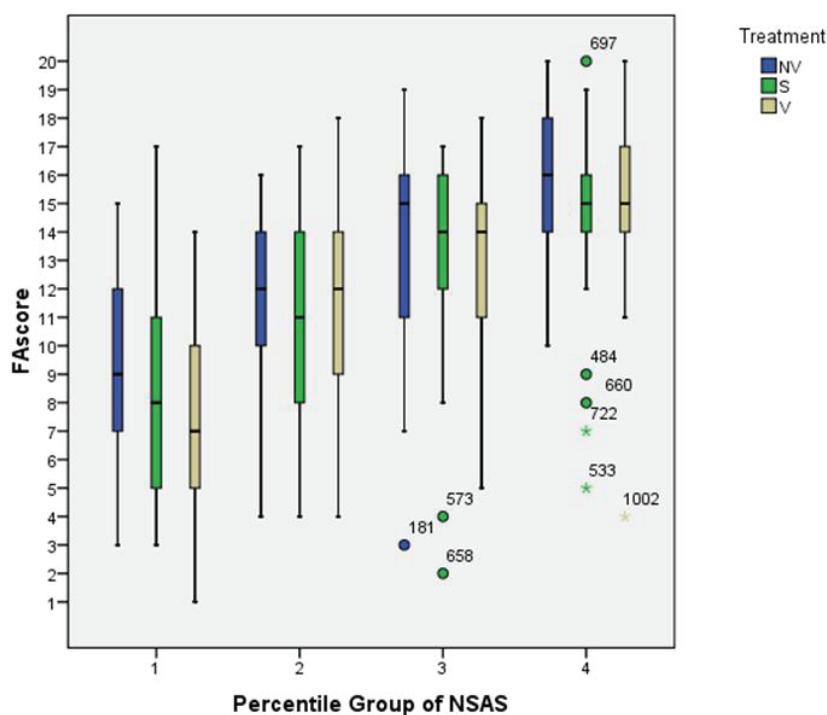


Figure 4.7 Conditioning on CogAT 6 scores

Six of these students were assigned to the standard treatment. What is particularly interesting is that three of the four students scoring exceptionally low in the 4th quartile standard treatment group were identified as ELL students. Although this result is suggestive at best, it may indicate that directions that work well enough for most students may fail to provide sufficient support to a small number of students, particularly ELL students. These students appeared greatly disadvantaged when receiving the standard directions with only two examples and very little explanation of the task. Another explanation is that they gained much more from the practice test than other students and were able to significantly improve their performance when they encountered the figure analogy format again on the CogAT.

Matching Strategy

In reviewing the item characteristics provided by IteMan, I observed that among low scoring examinees, the most popular response was by far the “matching” distractor. This is consistent with previous research (Gentile et al., 1977; Goldman et al., 1982; Vodegel-Matzen et al., 1994). Low scoring students overwhelmingly chose distractors that were identical to the C term in the analogy (A is to B as C is to ?) and sometimes even a distractor that was identical to the B term. In my final follow-up research question, I asked—*Do the verbal-dynamic or nonverbal-dynamic treatments reduce the popularity of matching distractors on non-matching items?* Out of 20 items, 11 offered distractor choices that matched either the B or C term (excluding the first three items where this was the answer). These items were #4-7 and #12-18. These items were not the most difficult, but comprised the middle range of items. Of the first-grade students, 35% chose the matching distractors on six or more of these items. For second-grade students, only 17% frequently chose matching distractors.

This observation appeared to vary across treatments in the first-grade sample where the proportion of students fitting this category ranged from 31% for nonverbal-

dynamic, to 36% for standard, to 39% for the verbal-dynamic group. Although these items showed mixed degrees of DIF in Table 4.7, a group of seven items that had distractors that clearly matched the C term of the analogy showed large DBF favoring the standard group over the verbal-dynamic group ($\beta = .23$) and small DBF for the standard over the nonverbal-dynamic group ($\beta = .08$). These analyses indicate that discouraging a matching strategy is especially important at first grade, but it is not clear that either of the innovative dynamic directions were effective in doing so.

Teacher Comments

The cover page for the test booklets asked the teachers to comment on any problems they had with test administration or any suggestions they had for improving the directions. Out of 46 classrooms, 10 teachers reported that their students had trouble focusing on the test with some indicating that the length of the test or the directions was a problem for their students. Seven teachers (all of whom had at least one ELL student in their classroom) suggested that the Spanish-English combination directions were confusing for their students and led to poor focus.

Summary

The results showed large mean differences between grade levels and ELL groups. A mean effect of grade was expected as the cognitive abilities measured by this test develop quickly at this age. The magnitude of differences in ELL groups did not appear to vary with treatment assignment, indicating that the test directions were not effective in reducing the differences in performance across ELL groups. The effect of treatment appeared to vary by grade level as second-grade students assigned to the nonverbal-dynamic treatment performed slightly better than second-grade students assigned to other treatments. This effect was localized partly to the double-rule items, which appeared slightly easier for this group in the DIF/DBF analyses.

Surprisingly, the verbal-dynamic treatment had no main effect relative to the standard treatment, though the treatment did have narrow effects on performance. First, students in this treatment group showed a slight advantage on test items in the middle range of difficulty that used addition or size-change rules. Second, there was an increasing slope across quartiles of performance whereby more able students appeared to gain the most from the verbal-dynamic treatment compared to lower performing students. This effect was especially clear in first grade. Neither of the dynamic-video treatments appeared to positively affect the number of students choosing matching distractors.

Variation in study procedures also appeared to have an effect. Classrooms that were not able to use the DVD/VHS videos with the directions actually performed better than classrooms that did use the videos. Possible explanations for this effect are explored in Chapter 5.

CHAPTER V DISCUSSION

The figure analogies test used in this study had not been used in other research. Therefore, generalizing the results first required demonstrating the psychometric adequacy of the test. Overall, the test showed excellent psychometric qualities and thus was appropriate to address the research questions that were posed in this study. The items showed a good range of difficulty and strong discrimination with a median point-biserial correlation of .50. The distributions of scores were also consistent with the test having sufficient ceiling and floors for the sample. Total scores on the test showed a strong correlation with CogAT 6 Nonverbal Battery scores ($r = 0.65$) and lower correlations with the CogAT Verbal Battery ($r = .43$), thereby demonstrating strong convergent and discriminant validity. The observation of significant growth across grade levels (.76 SD) is also consistent with a good measure of cognitive abilities.

It was not surprising that the test showed significantly lower scores for ELL students of about 2 points (.41 SD). Previous research has shown that the differences between ELL and non-ELL students can be quite large even on nonverbal tests. Unfortunately, the persistence of the mean differences across treatment groups suggests that the directions that were used did not go far enough in helping ELL students understand and perform better on the task. However, given the small ELL sample, it is possible that I was simply unable to detect a difference due to lack of power.

The Positive Effect of the Nonverbal Treatment

The primary difference between the standard and nonverbal-dynamic treatment was the provision of additional practice with feedback. It is unsurprising that this would create a small advantage in the latter group as previous research has shown that additional practice with feedback is beneficial, particularly when the total number of practice items is small (i.e., there is a diminishing return of additional practice items; Shute, 2007). The

observation of a treatment effect only at second grade is difficult to explain. I anticipated that if there were a difference across grades, then first-grade students would benefit more because they are less familiar with tests. However, in this school district, the second-grade students had not yet taken a formal test and were therefore quite test-naïve as well. Perhaps the second-grade students were still sufficiently unfamiliar with tests to benefit from the extra practice and possibly more motivated by the upcoming CogAT assessment.

One specific benefit of the nonverbal-dynamic treatment appeared to be an improvement in performance on the double-rule items. These items tend to be quite difficult for young students (Budoff et al., 1974). Significant differential bundle functioning (DBF) was observed on these items for students in the nonverbal-dynamic treatment compared to students in the standard treatment. This suggests that providing a more representative range of examples in the directions improves performance on similar items.

Narrow Effect of Verbal-Dynamic Directions

The verbal-dynamic directions were expected to lead to superior performance compared to the other directions because they provided more information about the correct answer and more guidance for strategy development. There are several potential explanations for why the verbal-dynamic directions did not affect overall performance. First, the addition of more instruction and discussion may have bored the more able students. Part of the directions included extensive discussion of why all three distractors are wrong. This may have made the directions too long as even options that are rarely selected must be discussed and students who chose other options (especially the correct option) may become bored (Shute, 2007). In pilot testing, it was clear that protracted discussion of the test without opportunities for the student to mark answers or respond led to more counterproductive behaviors. Adapting the amount of explanation provided to the

needs of students would avoid these effects, but this is difficult when the test is group administered.

Second, the increased verbal load of the verbal-dynamic treatment, though grounded in a strong research base and concretized by the video animations, could have introduced undesirable conditions for many students. More language load is clearly undesirable for ELL students when the language is not clear, relevant, and supported by visuals. However, increased language load could also have a negative impact on low-ability non-ELL students by increasing the cognitive load of the directions.

Third, the directions may have included too many teaching elements in a short time. Using a constructive solution strategy prior to examining the options, preferring reflective over impulsive solutions, and systematically eliminating options are all beneficial strategies for students who use them. However, students who do not already use these strategies may not acquire them without extensive practice. The observation that high-ability students seemed to gain more from the verbal-dynamic treatment is consistent with this explanation. In general, high-ability students are more likely to acquire strategies with relatively little training. Instead of going through these strategies quickly during the directions, more extensive practice activities that gradually introduce and reinforce these strategies might be more effective for less-able students.

Negative Effect of DVD

Perhaps the most surprising finding was the negative effect of using the DVD/VHS video in the directions, particularly in second grade. This unintentional, but serendipitous, variation was almost perfectly crossed with treatment assignment. Explanations for this finding are speculative but potentially warrant further study. One explanation is that children are more attentive to a teacher reading a script than to a video. This is consistent with my observation during pilot studies that I often needed to repeat part of the directions before students would act on directions provided by the

video. That is, if the video asked students to draw their answer, they often did not move to draw on the page until I prompted them to do so. A related explanation is that teachers are better able to manage students' attention when they are reading aloud—adjusting their pace, repeating sections, or re-engaging students as needed. The familiarity of the teacher's voice might also play an important role in helping students understand the directions. Any of these explanations would have important implications for computerized testing where students must manage their own attention and listen to unfamiliar voices.

Another potential explanation for the negative effect of DVD/VHS use is the English- Spanish audio combination. The use of combined English plus Spanish directions in all classrooms was intended to be a Universal Design feature that would support comprehension of Spanish-speaking ELL students in every classroom without harming the comprehension of non-ELL students. However, several teachers observed that the mix of directions caused a lack of focus during the Spanish portions. Although published tests increasingly provide Spanish translations, perhaps the intermingling of the two languages as well as the use of a single reader for both languages (so that it was less obvious when the English started again) was particularly confusing to young students. It is unclear how best to administer the combined directions since waiting to provide the Spanish sentences all at once following the English directions would probably be more confusing to ELL students. Additional research on the optimal method for providing translated directions during a group-administered test is needed, particularly in the context of developing test accommodations.

Implications for Designing Test Directions

Based on the pilot studies and full experiment, I reached five conclusions about the ideal design of test directions for young students. These conclusions could apply to items from ability or achievement domains, especially when innovative item formats are

being introduced. First, it was clear that good test directions should engage young students immediately in guessing an answer rather than expecting them to listen to extensive directions, particularly when the first example is quite simple. Students seemed to maintain their attention better when listening passages were short and interspersed with multiple opportunities to respond to the practice items.

Second, providing more examples helps some students improve their performance. Providing more examples offloads the need to explain an abstract task in detail. This is consistent with research that shows that schemes are better built from multiple examples than from long explanations (Anderson et al., 1984; Morrisett & Hovland, 1959). Additional examples also provide a more representative sample of test items. In reviewing existing tests, I observed that many test directions offered only basic example items that resembled the first few operational items. Such examples do not push students to develop a strategy that could be applied to a wide range of items.

Third, although I did not confirm my hypothesis that encouraging a constructive strategy would lead to better performance in the verbal-dynamic treatment, I did confirm that having students draw their answers on early items was a natural and effective method to teach a constructive strategy. Even students in the nonverbal-dynamic and standard treatment groups occasionally drew in answers spontaneously, indicating that this might be a preferred mode of responding. Having students draw their answers is thus a good method for encouraging constructive responses during practice, even though it may not be a practical mode of responding for the full test.

Fourth, multiple tactics are needed within the directions to slow students down and to encourage them to consider all of the response options before selecting an answer. In pilot testing, it appeared that many students completed the practice items without really understanding the task. Passively receiving feedback may not sufficiently challenge erroneous conceptions of an analogy that lead students to persist in seeking a matching solution over a true analogical solution. Even quite able students sometimes

impulsively chose matching distractors without considering all of the options. In the pilot study, I learned that students often could answer the questions correctly when they were encouraged to examine all options before picking one or when they were challenged to explain why they chose a particular option. In the verbal-dynamic directions, I tried to introduce more challenges to students' schemes to discourage both impulsivity and matching strategies, but effecting this change in a full class of students may require more individual adaption than group-administered directions following a standardized script can provide.

Finally, I found that using videos to provide directions could be problematic for young students. Even using relevant and interesting animations did not seem to hold students' attention compared to directions read aloud by a teacher. Future attempts to create computerized tests or to use video-based directions should take into account that test directions may be less effective when administered by a pre-recorded video than when read by teachers who can redirect students' attention and adjust the pace and content of the directions as needed.

No Directions at All?

As was mentioned in Chapter 3, teachers in two of the second-grade classrooms skipped the directions entirely so that students did not receive any feedback on performance. An intriguing observation is that the average score in both classes was 11 points, just below the average for the other second-grade classes. This confirms what I observed in pilot testing—that some students already seemed to know or easily learn what an analogy was based just on the items. When these students were simply given an opportunity to attempt the items, some answered the items correctly without any feedback. Although it is possible that these students had encountered analogies before, it appears that the first few test items (known as “teaching items”) are effective in guiding some students to an analogical solution strategy.

Adapting Instructions to Students

Part of the CogAT tradition is the “Betty Hagen rule”: Do not start the test until you are sure every child understands what he or she is supposed to do. When students attempt the first test item, all students should be on an even footing in their understanding of the demands of the task. Getting every student up to this criterion requires that the student who is least familiar with the task receive enough description and practice to understand the task and to answer the easiest items. Initially, I believed that introducing additional components to directions could have only a positive effect on test-naïve students and would have no net effect on test-wise students. However, this was not the case. Creating standardized directions that cater to the least able and least familiar student leaves the more prepared or more able students fidgeting through long directions that they do not need. Indeed, students and whole classrooms vary in their familiarity with tests and in how extensive the directions need to be. Therefore, a single set of directions that must be followed rigidly cannot satisfy the needs of students at both extremes of familiarity.

In addition to the challenge of differences in familiarity, this study also made apparent the important role of teachers in test administration. Teachers play a crucial role in adapting directions in small ways to help their students understand the test task. In the verbal-dynamic treatment, I attempted to add components that would standardize these adaptations, including hints and guidance, but teachers still played a crucial role in responding to student behaviors and encouraging their class to engage with the video directions. Thus, it became clear through this study that any improvements in directions or practice would be only as effective as the teacher using them and only to the extent that the teacher adapts the directions to the particular characteristics of his/her students. One option for improving directions is to develop brief training materials that help teachers understand the types and extent of modifications that can be made. Another

alternative is to make additional practice items with brief feedback available as needed. The teacher could then adapt the directions without seriously harming standardization.

Another solution, which might be even more effective, is to develop computer-based adaptive directions that could provide the needed flexibility in practice and feedback. Computer-based directions can respond more adaptively to what students can already do and provide more or less practice as needed on an individual basis. They can also permit the sort of trial-and-solution strategy that students commonly use when learning new computer games. Computer-based directions could also target the feedback to provide only useful (and brief) corrections for errors made on practice items and support this feedback with appropriate visual demonstrations. A further benefit of computer-based practice would be that examinees could elect to hear the directions in one of several languages. However, given the observations in this study, especially the pilot tests, encouragement and clarification by a live instructor may be important to support computerized practice. In fact, the need for intensive proctoring may be a serious limitation to computer-based testing for young students.

Choice of practice and teaching items

Another clear finding consistent with several other research studies is that some young students develop a “matching” strategy on figure analogies whereby they consistently choose a distractor that matches one of the terms in the analogy stem rather than the distractor that fits a true analogical relationship. This was especially clear among first-grade students where 35% chose the matching distractor on six or more of the 11 items with such distractor choices. Students who pursue this unproductive strategy apparently do not have a clear idea of what a true analogical relationship is even after completing the practice items. On the one hand, this may be related to the construct of interest (i.e., that lower-ability students in the primary grades are less likely to reason analogically). On the other hand, some higher-ability students may be pursuing this

strategy throughout the test due to a misunderstanding of the task that is abetted by simplistic practice items (see discussion of superordinate terms in verbal classification tasks in Chapter 2).

This finding leads me to question the use of any items with a matching solution for the practice items and early test items. It is quite possible that this strategy is encouraged in middle- to low-ability students by the use of such items early in the task when students are developing their understanding of analogies and are seeking to relate the task to conceptual schemes they may already have. Anecdotally, I noted that a large number of students in every treatment group initially marked the matching answer choice on the second practice item on the test before erasing and changing their answer to the correct non-matching option. I attempted to discourage such a solution by using three non-matching practice items in the verbal-dynamic and nonverbal-dynamic treatments, but this was not successful in discouraging the strategy given that a similar number of students in each group pursued the matching strategy.

Limitations

The relatively small sample size for the ELL students (N=126) limited the power of the analyses to detect the effects of directions in reducing mean differences between ELL and non-ELL students. However, the information available indicates that the directions were not effective. If this is confirmed in larger samples, it indicates that small differences in familiarity with the test task cannot explain the large score gaps observed between ELL and non-ELL students. If this is so, the next step would be to investigate more extensive practice activities that could have a greater impact on students' understanding of an analogy as well as provide a better chance for students to learn and adopt useful strategies. If extensive practice also does not narrow mean differences, then one might conclude that the mean differences may result from large disparities in opportunity to learn that are present even for nonverbal reasoning tasks.

Another limitation of this study was that the verbal-dynamic treatment confounded several potentially beneficial features of test directions. This was intentional, because I wanted to create an optimal set of directions to contrast with the other treatments. However, the failure to find a positive effect of the verbal-dynamic treatment does not mean that any of these features fail to benefit students. For example, previous research indicated that encouraging students to put the analogical rule into words (a verbalization strategy) could improve performance (Malloy et al., 1987). Teaching this strategy may require more explicit and lengthy instruction than was provided in the verbal-dynamic directions. Because of the confounding of test design elements, I could also not confirm that offloading orienting directions (e.g., “look at the top row”) through video animation improved performance for students. This feature especially seems unlikely to cause any harm and might reasonably be included in video- or computer-based directions in the future. Finally, although encouraging students to draw their answers on the early items appeared to be engaging, I was unable to determine if this activity alone would have any positive effect on test performance.

Future Research

Well-designed directions should provide the vast majority of examinees sufficient opportunities to learn the task. As in regular instruction, lower-ability examinees will require more direct instruction and practice before they will understand the task. Thus, test directions should provide adequate practice opportunities for these students. However, test developers must also be concerned with boring or disengaging bright examinees. Overall, the results indicate that rather than seeking to design the ideal set of directions, the goal should be to provide a testing environment that adapts to the needs of the particular students in a classroom without greatly compromising the standardization of the test. Future research might explore the role that computer-based directions and advanced practice activities could play in providing this adaptable test preparation.

REFERENCES

- Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment, 8*(3), 231-257.
- Abedi, J. (2006). Language issues in item development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377-398). Mahwah, NJ: Lawrence Erlbaum Associates.
- Abedi, J., Kao, J. C., Leon, S., Sullivan, L., Herman, J., Pope, R., Nambiar, V., & Mastergeorge, A. M. (2009). *Exploring factors that affect the accessibility of reading comprehension assessments for students with disabilities: A study of segmented text*. Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.
- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Federation of Teachers. (2004). *Closing the achievement gap: Focus on Latino students (Policy Brief Number 17)*. Washington, D.C.: American Federation of Teachers. (ERIC Document Service No. ED497878)
- Anastasi, A. (1937). *Differential psychology: Individual and group differences in behavior*. New York: The Macmillan Company.
- Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36*, 1086-1093.
- Anderson, J. R., Farrell, R., & Sauers, R. (1984). Learning to program in LISP. *Cognitive Science, 8*(2), 87-129.
- Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology Learning Memory and Cognition, 23*, 932-945.
- Baker, E. L., Atwood, N. K., & Duffy, T. M. (1988). Cognitive approaches to assessing the readability of text. In A. Davison, & G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered* (pp. 55-83). Hillsdale, NJ: Lawrence Erlbaum.
- Bergman, I. B. (1980). The effects of test-taking instruction vs. practice without instruction on the examination responses of college freshmen to multiple-choice, open-ended, and cloze type questions. *Dissertation Abstracts International: Section A. Humanities and Social Sciences, 41*(1-A), 176.
- Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence, 8*(3), 205-238.

- Biesheuvel, S. (1972). Adaptability: Its measurement and determinants. In L. J. Cronbach, & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptation* (pp. 47-62). The Hague: Mouton.
- Binet, A., & Simon, T. (1908/1961). The development of intelligence in the child. In J. J. Jenkins, & D. G. Paterson (Eds.), *Studies in individual differences* (pp. 81-111). New York: Appleton-Century-Crofts.
- Bond, L. (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 429-444). New York, NY: American Council on Education/Macmillan.
- Bracken, B. A. (2007). Creating the optimal preschool testing situation. In B. A. Bracken, & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 137-154). Mahwah, NJ: Lawrence Erlbaum Associates.
- Bracken, B. A., & McCallum, R. S. (1998). *Universal Nonverbal Intelligence Test (UNIT)*. Itasca, IL: Riverside.
- Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.
- Bridgeman, B., & Buttram, J. (1975). Race differences on nonverbal analogy test performance as a function of verbal strategy training. *Journal of Educational Psychology*, 67(4), 586-590.
- Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS: 88. *Chance*, 14(1), 10-21.
- Brooks-Gunn, J., & Markman, L. B. (2005). The contribution of parenting to ethnic and racial gaps in school readiness. *The Future of Children*, 15(1), 139-168.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.
- Budoff, M., Gimon, A., & Corman, L. (1974). Learning potential measurement with Spanish-speaking youth as an alternative to IQ tests: A first report. *Interamerican Journal of Psychology*, 8, 233-246.
- Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (CSE Technical Report No. 448). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Callahan, C. M. (2005). Identifying Gifted Students From Underrepresented Populations. *Theory into Practice*, 44(2), 98-104.
- Carlstedt, B., Gustafsson, J. E., & Ullstadius, E. (2000). Item sequencing effects on the measurement of fluid intelligence. *Intelligence*, 28(2), 145-160.
- Carman, C. A., & Taylor, D. K. (2009). Socioeconomic status effects on using the Naglieri Nonverbal Ability Test (NNAT) to identify the gifted/talented. *Gifted Child Quarterly*. Advance online publication. doi:10.1177/0016986209355976

- Cathers-Schiffman, T. A., & Thompson, M. S. (2007). Assessment of English-and Spanish-speaking students with the WISC-III and Leiter-R. *Journal of Psychoeducational Assessment*, 25(1), 41-52.
- Chrispeels, J. H., & Rivero, E. (2001). Engaging Latino families for student success: How parent education can reshape parents' sense of place in the education of their children. *Peabody Journal of Education*, 76(2), 119-169.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17(1), 31-44.
- Clemans, W. V. (1971). Test administration. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 188-201). Washington, D.C.: American Council on Education.
- Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355-386). New York, NY: American Council on Education/Macmillan.
- Corman, L., & Budoff, M. (1973). A comparison of group and individual training procedures on the raven learning potential measure with Black and White special class students. *Studies in Learning Potential*, 3(57), 1-16.
- Crocker, L. (2005). Teaching for the test: How and why test preparation is appropriate. In R. P. Phelps, E. D. Roeber, D. Rogosa, & J. Fremer (Eds.), *Defending Standardized Testing* (pp. 159-174). Mahwah, NJ: Lawrence Erlbaum Associates.
- Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York, NY: Harper and Row.
- Crowe, D. E. (1982). The use of practice programs to improve test scores of elementary school students. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 42(07), 3116.
- Cummins, J. (1980). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL Quarterly*, 14(2), 175-187.
- Dague, P. (1972). Development, application and interpretation of tests for use in French-speaking black Africa and Madagascar. In L. J. Cronbach, & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptation* (pp. 63-74). The Hague: Mouton.
- Davison, A., & Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, 17, 187-209.
- DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics*, 34, 149-170.
- Detterman, D. K., & Andrist, C. G. (1990). Effect of instructions on elementary cognitive tasks sensitive to individual differences. *American Journal of Psychology*, 103(3), 367-401.

- Dreisbach, M., & Keogh, B. K. (1982). Testwiseness as a factor in readiness test performance of young Mexican-American children. *Journal of Educational Psychology, 74*(2), 224-229.
- Evans, F. R., & Pike, L. W. (1973). The effects of instruction for three mathematics item formats. *Journal of Educational Measurement, 10*(4), 257-271.
- Feist, M. I., & Gentner, D. (2007). Spatial language influences memory for spatial scenes. *Memory and Cognition, 35*(2), 283-296.
- Ferguson, G. A. (1954). On learning and human ability. *Canadian Journal of Psychology, 8*(2), 95-112.
- Flanagan, D. P., Kaminer, T., Alfonso, V. C., & Rader, D. E. (1995). Incidence of basic concepts in the directions of new and recently revised American intelligence tests for preschool children. *School Psychology International, 16*(4), 345-364.
- Frisby, C. L. (1999). Culture and test session behavior: Part I. *School Psychology Quarterly, 14*, 263-280.
- Gaster, L., & Clark, C. (1995). *A guide to providing alternate formats*. West Columbia, SC: Center for Rehabilitation Technology Services. Retrieved from ERIC Data base. (ED405689)
- Gentile, J. R., Tedesco-Stratton, L., Davis, E., Lund, N. J., & Agunanne, B. C. (1977). Associative responding versus analogical reasoning by children. *Intelligence, 1*(4), 369-380.
- Gentner, D. (1978). On relational meaning: The acquisition of verb meaning. *Child Development, 49*(4), 988-998.
- Gentner, D. (2003). Why we're so smart. In D. Gentner, & S. Goldin-Meadow (Eds.), *Language in mind* (pp. 195-235)
- Glutting, J. J., & McDermott, P. A. (1989). Using "teaching items" on ability tests: A nice idea, but does it work? *Educational and Psychological Measurement, 49*, 257-268.
- Glutting, J. J., Oakland, T., & Konold, T. R. (1994). Criterion-related bias with the Guide to the Assessment of Test-Session Behavior for the WISC-III and WIAT: Possible race/ethnicity, gender, and SES effects. *Journal of School Psychology, 32*(4), 355-369.
- Goldman, S. R., Pellegrino, J. W., Parseghian, P., & Sallis, R. (1982). Developmental and individual differences in verbal analogical reasoning. *Child Development, 53*(2), 550-559.
- Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. *Child Development, 62*(1), 1-22.
- Greenfield, P. M., Quiroz, B., & Raeff, C. (2000). Cross-cultural conflict and harmony in the social construction of the child. *New Directions for Child and Adolescent Development, 2000* (87), 93-108.

- Hammill, D. D., Pearson, N. A., & Widerholdt, J. L. (1996). *Comprehensive Test of Nonverbal Intelligence*. Austin, TX: PRO-ED.
- Harris, B., Rapp, K. E., Martinez, R. S., & Plucker, J. A. (2007). Identifying English language learners for gifted and talented programs: Current practices and recommendations for improvement. *Roeper Review*, 29, 26-29.
- Hart, T. R., & Risley, B. M. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H Brookes.
- Haslerud, G. M., & Meyers, S. (1958). The transfer value of given and individually derived principles. *Journal of Educational Psychology*, 49(6), 293-298.
- Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms*. New York, NY: Cambridge University Press.
- Heck, R. H. (2001). Multilevel modeling With SEM. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 89-127). Mahwah, NJ: Lawrence Erlbaum Associates.
- Heck, R. H., & Thomas, S. L. (2000). *An introduction to multilevel modeling techniques*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hess, R. D., & Shipman, V. C. (1965). Early experience and the socialization of cognitive modes in children. *Child Development*, 36(4), 869-886.
- Hessels, M. G. P., & Hamers, J. H. M. (1993). The learning potential test for ethnic minorities. In J. H. M. Hamers, K. Sijtsma & Ruijssenaars, A. J. J. M. (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 285-313). Amsterdam: Swets & Zeitlinger.
- Humphreys, L. G. (1976). Strategy training has no significant effect on race differences in nonverbal reasoning. *Journal of Educational Psychology*, 68(2), 128-129.
- ITEMAN (Version 3.6) [Computer Software]. (2006). St. Paul, MN: Assessment Systems Corporation.
- Jacobs, P. I., & Vandeventer, M. (1971). The learning and transfer of double-classification skills by first graders. *Society for Research in Child Development*, 42, 149-159.
- James, W. S. (1953). Symposium on the effects of coaching and practice in intelligence tests: Coaching for all recommended. *British Journal of Educational Psychology*, 23, 155-162.
- Jencks, C. (1998). Racial bias in testing. In C. Jencks, & M. Phillips (Eds.), *The black-white test score gap* (pp. 55-85). Washington, D.C.: Brookings Institution.
- Johnstone, C. J., Thompson, S. J., Bottsford-Miller, N. A., & Thurlow, M. L. (2008). Universal design and multimethod approaches to item review. *Educational Measurement: Issues and Practice*, 27(1), 25-36.
- Judd, C. H. (1908). The relation of special training to general intelligence. *Educational Review*, 36(28), 42.

- Kaufman, A. S. (1978). The importance of basic concepts in individual assessment of preschool children. *Journal of School Psychology, 16*, 207-211.
- Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children-2 (KABC-2)*. Circle Pines, MN: American Guidance Service.
- Ketterlin-Geller, L. R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice, 27*(3), 3-16.
- Kittell, J. E. (1957). An experimental study of the effect of external direction during learning on transfer and retention of principles. *Journal of Educational Psychology, 48*(6), 391-405.
- Klausmeier, H. J., & Feldman, K. V. (1975). Effects of a definition and a varying number of examples and nonexamples on concept attainment. *Journal of Educational Psychology, 67*(2), 174-178.
- Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers. Retrieved from ERIC Data base. (ED454703)
- Kopriva, R. J. (2008). *Improving testing for English language learners*. New York, NY: Routledge.
- Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*, 435-447.
- Lakin, J. M., & Lai, E. R. (2010, April). *Exploring reliability and decomposing error variance of verbal, quantitative, and nonverbal ability for English-language learner students*. Paper presented at the annual meeting of the American Educational Researcher Association, Denver, CO.
- Lakin, J. M., & Lohman, D. F. (2009, July). *Predictive validity of ability tests for academic achievement similar for English-language learners and non-ELL students*. Poster presented at the biannual meeting of the International Society for the Study of Individual Differences, Chicago, IL.
- LeFevre, J. A., & Dixon, P. (1986). Do written instructions need examples? *Cognition and Instruction, 3*(1), 1-30.
- Leighton, J. P. (2004). The assessment of logical reasoning. In R. J. Sternberg, & J. P. Leighton (Eds.), *The nature of reasoning* (pp. 291-312). New York, NY: Cambridge University Press.
- Lewis, J. D. (2001). *Language isn't needed: Nonverbal assessments and gifted learners*. Paper presented at Growing Partnerships for Rural Special Education, San Diego, CA. Retrieved from ERIC Data base. (ED453026)
- Lindquist, E. F. (Ed.). (1951). *Educational measurement*. Washington, DC: American Council on Education.

- Linn, R. L. (Ed.). (1989). *Educational measurement*. New York, NY: American Council on Education/Macmillan.
- Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, *50*(4), 315-353.
doi:10.1016/j.cogpsych.2004.09.004
- Lohman, D. F. (2000). Complex information processing and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (2nd ed., pp. 285-340). New York, NY: Cambridge University Press.
- Lohman, D. F. (2001). Fluid intelligence, inductive reasoning, and working memory: Where the theory of Multiple Intelligences falls short. In N. Colangelo, & S. Assouline (Eds.), *Talent development IV: Proceedings from the 1998 Henry B. & Jocelyn Wallace National Research Symposium on talent development* (pp. 219-228). Scottsdale, AZ: Gifted Psychology Press.
- Lohman, D. F., & Al-Mahrzi, R. (2003). *Personal standard errors of measurement*. Unpublished manuscript.
- Lohman, D. F., & Hagen, E. P. (2001). *Cognitive Abilities Test (Form 6)*. Itasca, IL: Riverside Publishing Company.
- Lohman, D. F., & Hagen, E. P. (2002). *Cognitive Abilities Test (Form 6): Research handbook*. Itasca, IL: Riverside Publishing Company.
- Lohman, D. F. (in press). *Cognitive Abilities Test (Form 7)*. Itasca, IL: Riverside Publishing Company.
- Lohman, D. F., Korb, K. A., & Lakin, J. M. (2008). Identifying academically gifted English-language learners using nonverbal tests: A comparison of the Raven, NNAT, and CogAT. *Gifted Child Quarterly*, *52*(4), 275-296.
- Lubienski, S. T. (2000). A clash of social class cultures? Students experiences in a discussion-intensive seventh-grade mathematics classroom. *Elementary School Journal*, *100*, 377-403.
- Malloy, T., Mitchell, C., & Gordon, O. (1987). Training cognitive strategies underlying intelligent problem solving. *Perceptual and Motor Skills*, *64*, 1039-1046.
- Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431-467.
- Martiniello, M. (2003). *Institutional rules and regulations of public and publicly-subsidized private schools: A literature review*. Unpublished Ph.D., Harvard University,
- Martiniello, M. (2008). *Item characteristics predicting DIF for ELLs in math assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.
- Maspons, M. M., & Llabre, M. M. (1985). The influence of training Hispanics in test taking on the psychometric properties of a test. *Journal for Research in Mathematics Education*, *16*, 177-183.

- Mayer, R. E. (2001). *Multimedia learning*. New York, NY: Cambridge University Press.
- McCallum, R. S., Bracken, B. A., & Wasserman, J. (2001). *Essentials of nonverbal assessment*. New York, NY: Wiley.
- Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-355). New York, NY: American Council on Education/Macmillan.
- Morrisett, L., & Hovland, C. I. (1959). A comparison of three varieties of training in human problem solving. *Journal of Experimental Psychology*, 58(1), 52-55.
- Muthén, L.K. (1999, October 29). Intraclass correlations [Msg 2 & 4]. Message posted to <http://www.statmodel.com/discussion/messages/12/18.html?1253640879>
- Naglieri, J. A. (1996). *Naglieri Nonverbal Ability Test*. San Antonio, TX: Harcourt Brace Educational Measurement.
- National Assessment Governing Board (n.d.). *NAEP Data Explorer [Data file]*. Retrieved July 1, 2009, from <http://nces.ed.gov/nationsreportcard/ltddata/>
- Neuhaus, M. (1967). Modifications in the administration of the WISC performance subtests for children with profound hearing losses. *Exceptional Children*, 33, 573-574.
- Oakland, T., & Harris, J. G. (2009). Impact of test-taking behaviors on full-scale IQ scores from the Wechsler Intelligence Scale for Children-IV, Spanish Edition. *Journal of Psychoeducational Assessment*, 27(5), 366-373.
- Oller, J. W., Jr., Kim, K., & Choe, Y. (2001). Can instructions to nonverbal IQ tests be given in pantomime? Additional applications of a general theory of signs. *Semiotica*, 133(1/4), 15-44.
- Ortar, G. R. (1960). Improving test validity by coaching. *Educational Research*, 2(2), 137-142.
- Ortar, G. (1972). Some principles for adaptation of psychological tests. In L. J. Cronbach, & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptation* (pp. 111-120). The Hague: Mouton.
- Ortiz, S. O., & Ochoa, S. H. (2005). Conceptual measurement and methodological issues in cognitive assessment of culturally and linguistically diverse individuals. In R. L. Rhodes, S. H. Ochoa & S. O. Ortiz (Eds.), *Assessing culturally and linguistically diverse students: A practical guide* (pp. 153-167). New York: Guilford Press.
- Palmer, D. J., Olivarez Jr., A., Willson, V. L., & Fordyce, T. (1989). Ethnicity and language dominance: Influence on the prediction of achievement based on intelligence test scores in nonreferred and referred samples. *Learning Disability Quarterly*, 12(4), 261-274.
- Peña, E. D., & Quinn, R. (1997). Task familiarity: Effects on the test performance of Puerto Rican and African American children. *Language, Speech, and Hearing Services in Schools*, 28(4), 323-332.

- Pitoniak, M. J., Young, J. W., Martiniello, M., King, T. C., Buteux, A., & Ginsburgh, M. (2009). *Guidelines for the assessment of English-language learners*. Princeton, NJ: ETS.
- Powers, S., & Barkan, J. H. (1986). Concurrent validity of the Standard Progressive Matrices for Hispanic and non-Hispanic seventh-grade students. *Psychology in the Schools, 23*, 333-336.
- Raudenbush, S., Bryk, A., Cheong, Y.F., Congdon, R. (2004). *HLM 6.0: Hierarchical linear and nonlinear modeling* [computer program]. Lincolnwood, IL: Scientific Software International.
- Reese, L., Balzano, S., Gallimore, R., & Goldenberg, C. (1995). The concept of educación: Latino family values and American schooling. *International Journal of Educational Research, 23*(1), 57-81.
- Resing, W.C.M., Tunteler, E., de Jong, F.M., & Bosma, T. (2009). Dynamic testing in indigenous and ethnic minority children. *Learning and Individual Differences, 19*, 445-450.
- Rivera, C., & Collum, E. (2006). *State assessment policy and practice for English language learners: A national perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Rivera, C., & Stansfield, C. W. (2003). The effects of linguistic simplification of science test items on score comparability. *Educational Assessment, 9*(3&4), 79-105.
- Rivera, C., Stansfield, C. W., Scialdone, L., & Sharkey, M. (2000). *An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998-1999*. Arlington, VA: The George Washington University, Center for Equity and Excellence in Education. Retrieved from ERIC Data base. (ED445037)
- Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 42-55.
- Scarr, S. (1978). From evolution to Larry P., or what shall we do about IQ tests? *Intelligence, 2*(4), 325-342.
- Scarr, S. (1981). Implicit messages: A review of "Bias in mental testing". *American Journal of Education, 89*(3), 330-338.
- Schnotz, W. K., & Kurschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review, 19*, 469-508.
- Schwarz, P. A. (1971). Prediction instruments for educational outcomes. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 303-334). Washington, D.C.: American Council on Education.
- Shute, V. J. (2007). Focus on formative feedback (Research report No. RR-07-11). Retrieved from Educational Testing Service website: <http://www.ets.org/research/researcher/RR-07-11.html>

- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage Publications.
- Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist*, 27(1), 5-32.
- Snow, R. E., & Lohman, D. F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology*, 76(3), 347-376.
- Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English Language Learners. *Educational Researcher*, 37(4), 189-199.
- Spybrook, J., Raudenbush, S. W., Liu, X., & Martinez, A. (2008). *Optimal Design Version 1.76 [Computer Software and Manual]*.
- Stanley, J.C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356-442). Washington, D.C.: American Council on Education.
- Stenning, K., & Monaghan, P. (2004). Strategies and knowledge representation. In R. J. Sternberg, & J. P. Leighton (Eds.), *The nature of reasoning* (pp. 129-168). New York, NY: Cambridge University Press.
- Stern, W. (1914). The psychological methods of testing intelligence. In G.M. Whipple (Ed. & Trans.), *Educational psychology monographs* (No. 13). Baltimore: Warwick & York. Retrieved from <http://hdl.handle.net/2027/mdp.39015014498391>
- Sternberg, R. J., & Nigro, G. (1980). Developmental patterns in the solution of verbal analogies. *Child Development*, 51(1), 27-38.
- Sullivan, A. M. (1964). *The relation between intelligence and transfer* (Doctoral thesis, McGill University, Montreal, Quebec).
- Sullivan, A. M., & Skanes, G. R. (1971). Differential transfer of training in bright and dull subjects of the same mental age. *British Journal of Educational Psychology*, 41(3), 287-293.
- Sullivan, P. M. (1982). Administration modifications on the WISC-R performance scale with different categories of deaf children. *American Annals of the Deaf*, 127(780), 788.
- Thomas, W. P., & Collier, V. (1997). *School Effectiveness for Language Minority Students* No. NCBE Resource Collection Series No. 9). Washington, D.C.: National Clearinghouse for Bilingual Education, George Washington University, Center for the Study of Language and Education.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal Design Applied to Large Scale Assessments* (NCEO Synthesis Report No. 44) National Center on Educational Outcome.
- Thompson, S. J., Thurlow, M., & Malouf, D. B. (2004). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology*, 6(1), 1-15.

- Thorndike, E. L. (1914). *Educational psychology: A briefer course*. New York, NY: Teachers College, Columbia University.
- Thorndike, E. L. (1922). Practice Effects in Intelligence Tests. *Journal of Experimental Psychology*, 5(2), 101.
- Thorndike, E. L. (1927). The law of effect. *American Journal of Psychology*, 39(212), 222.
- Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.
- Thorndike, R. M., & Hagen, E. (1974). *Cognitive Abilities Test (Multilevel Edition): Technical manual*. Boston, MA: Houghton Mifflin.
- Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale: Fourth Edition*. Itasca, IL: Riverside Publishing.
- Traxler, A. E. (1951). Administering and scoring the objective test. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 329-416). Washington, DC: American Council on Education.
- Tunteler, E., Pronk, C., & Resing, W.C.M. (2008). Inter-and intra-individual variability in the process of change in the use of analogical strategies to solve geometric tasks in children: A microgenetic analysis. *Learning and Individual Differences*, 18, 44-60.
- van de Vijver, F., & Poortinga, Y. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable. *European Journal of Psychological Assessment*, 8(1), 17-24.
- Verhallen, M. J. A. J., Bus, A. G., & de Jong, M. T. (2006). The promise of multimedia stories for kindergarten children at risk. *Journal of Educational Psychology*, 98(2), 410-419.
- Vodegel-Matzen, L. B. L., Van der Molen, M. W., & Dudink, A. C. M. (1994). Error analysis of Raven test performance. *Personality and Individual Differences*, 16(3), 433-445.
- Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University.
- Walpole, M., McDonough, P. M., Bauer, C. J., Gibson, C., Kanyi, K., & Toliver, R. (2005). This test is unfair: Urban African American and Latino high school students' perceptions of standardized college admission tests. *Urban Education*, 40(3), 321-349.
- Wechsler, D. (2003). *Wechsler Intelligence Scale for Children 4th Edition (WISC-IV)*. San Antonio, TX: Harcourt Assessment.
- Wechsler, D. (2005). *Wechsler Intelligence Scale for Children-Fourth Edition Spanish*. San Antonio, TX: Harcourt Assessment.
- Weiss, L. G., Saklofske, D. H., Prifitera, A., & Holdnack, J. A. (2006). *WISC-IV advanced clinical interpretation*. Burlington, MA: Elsevier.

Whitely, S. E., & Dawis, R. V. (1974). Effects of cognitive intervention on latent ability measured from analogy items. *Journal of Educational Psychology, 66*, 710-717.

Yano, Y., Long, M. H., & Ross, S. (1994). The effects of simplified and elaborated texts on foreign language reading comprehension. *Language Learning, 44*(2), 189-219.

APPENDIX A

DIRECTIONS FOR ADMINISTERING STANDARD DIRECTIONS

Research Study Directions for Administration

**Cognitive Abilities Test
Practice Form
Grades 1-2**

Standard Directions

Introduction

The *Cognitive Abilities Test™ (CogAT®)* is an integrated series of tests that appraises the cognitive development of students in kindergarten through grade 12. It assesses learned reasoning abilities in three areas of cognitive processing that are essential for success in school—verbal, quantitative, and nonverbal (or figural) reasoning.

Riverside Publishing is conducting a research study in first and second grade to try out this practice test which uses new types of test directions for *CogAT*. This research study will help us develop new forms of test directions and practice for *CogAT* that will help all students, but especially English-language learners, to perform their best on the test. We appreciate your participation in this study.

Please study this manual in advance of the testing day so that there will be no hesitation on your part when the test is administered.

Following the administration procedures in this manual exactly will ease the task of administering the test and will help ensure that students do not receive too much or too little assistance from the test administrator.

You will be administering one practice test which will take about 30 minutes, but you may want to allow more time, in case it is needed.

Preparing for Testing

- There should be one copy of this *Directions* manual for you and one test booklet for each student. Please note that different classrooms will receive different forms of the test. If you need additional copies of the student booklets, make sure that any copies you borrow have the exact same cover as the rest in your class. Make sure that your *Directions* manual is appropriate for the student booklets (i.e., the “Standard” Directions must be used in a classroom with “Standard or Cartoon” booklets. Store all of the test materials in a secure place.
- Be sure that each child has a soft, black-lead (No. 2) pencil with a good eraser. Have some extra pencils on hand in case they are needed.
- **Prepare a place marker for each child.** Place markers may be rectangular pieces of cardboard or folded pieces of paper, about 2 inches by 4 inches. **Place markers are not included in the test materials and must be prepared by the test administrator or an aide prior to testing.**
- **The test directions require playing a DVD for the class.** Arrange to have a TV with a DVD player on the day of testing. **Watch the DVD ahead of time while following along in the test booklet so that you know what special requirements your set of directions have.**

Preparing student test booklets:

Student information is required on the cover of the test booklet. This information should be completed by the teacher. This information may be completed before or after testing, but must be complete before you return the materials to the testing coordinator.

- Student names are not being collected on the test booklets to protect student privacy. However, you may write the **first name of the student** on the cover so that they complete the right booklet.
- Instead of student names, record each student's **state ID number** on the front of the test booklet.
- Record the student's ELL status using the most recent English proficiency results.

During Testing

Follow the directions for administering the subtests **exactly** as they appear in this manual. You are provided with a full script of the DVD in green and orange print. Read aloud only the test directions printed in black. Directions in black within parentheses are for the test administrator and are not to be read to students. Follow along in the script throughout the directions as there are times when you may need to pause the video to give students a chance to mark answers or respond to questions.

Make sure that students understand the test directions and the sample questions before you begin each test. Answer any questions and help any students having difficulty. **Do not start the test until students understand what they are supposed to do.** Once the testing begins, you may not answer questions about specific items. However, questions about procedures and the mechanics of test taking may be answered at any time.

Move around the room while students are taking each test to make sure that they are working on the correct subtest and are filling in the answer bubbles appropriately.

There are no time limits for the subtests you will administer. Pace the students as they answer each test question. Allow enough time for all students to attempt to answer each question before telling students to go on to the next question. **Carefully record the exact time you begin and end each subtest.** Space to record this information is provided in this manual within the directions for each subtest.

In the space provided, please report any incident that could have an adverse effect on the test results.

After Testing

- Check that all students completed all items.
- Inspect each answer document and erase any stray marks.
- Complete “Teacher Information – Grade 1-2 Short Form” document as fully as possible
 - Look at the DVD or its envelope to find the **Version of DVD used** in your classroom (either Standard, Cartoon, Verbal Dynamic, or Nonverbal Dynamic).
- Stack the booklets with the “Teacher Information... Form” on top and bind with the rubber band provided.
- Return all test materials (including DVD) to the building test coordinator or other designated person.

Part 2: Administering the Test

Figure Matrices – Standard Directions

Today we will solve some puzzles. Some puzzles will be easy to solve. Others will be difficult to solve, but you should try your best. Let's try some together. *(start DVD)*

Open your book. You should be on the page with the **keys** across the top.

Abre tu libro. Debes de estar en la página con las **llaves** que atraviesan la parte de arriba.

Put your marker under the **broom**.

Pon tu cursor debajo de la **escoba**.

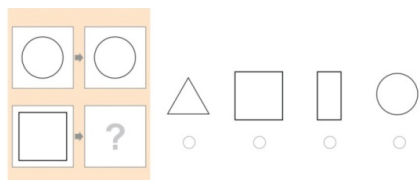
Now watch the video. *Ahora mira el video.*

Let's look at this puzzle together.

Vamos a mirar juntos este rompecabezas.

The large square is like a puzzle. This puzzle has a piece missing. You need to find the missing piece.

El cuadrado grande es como un rompecabezas. Le hace falta una pieza a este rompecabezas. Tienes que encontrar la pieza que falta.



The pictures on the right show possible pieces that finish the puzzle.

Los dibujos en el lado derecho muestran las piezas que posiblemente pueden completar el rompecabezas.

One of the pieces goes on the question mark to finish the puzzle.

Una de las piezas va sobre el signo de interrogación para completar el rompecabezas.

The second picture shows a square just like the square in the puzzle. This is the answer.

El segundo dibujo muestra un cuadrado igual al cuadrado que está en el rompecabezas. Esta es la respuesta correcta.

Now look at your book.

Ahora mira tu libro.

Fill in the bubble under the square to show it is the right answer.

Rellena el círculo debajo del cuadro para mostrar que es la respuesta correcta.

(Pause the DVD and make sure students have correctly filled in the answer)

Now we'll do another puzzle together. Put your marker under the **fish**.

Ahora haremos otro rompecabezas. Pon tu cursor debajo del **pescado**.

Now watch the video.

Ahora mira el video.

This puzzle has a piece missing. You need to find the missing piece.

Le hace falta una pieza a este rompecabezas. Tienes que encontrar la pieza que falta.

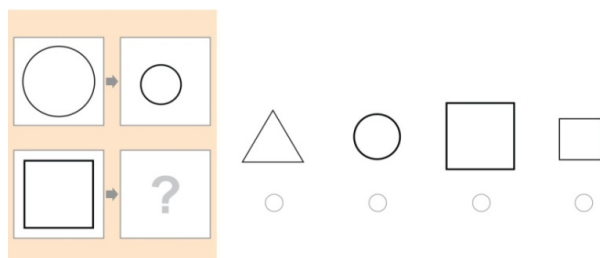
Different puzzles have different rules.

Cada rompecabezas tiene sus propias reglas.

Which is the missing piece for this puzzle?

¿Cuál es la pieza que hace falta para este rompecabezas?

(Long pause in audio. Pause DVD here if students need more time to mark their answer.)



The fourth picture shows a smaller square.

El cuarto dibujo es el cuadrado más chico.

This is the correct answer.

Esta es la respuesta correcta.

Now look at your book.

Ahora mira tu libro.

Fill in the bubble under the small square to show it is the right answer.

Rellena el círculo debajo del cuadrado más chico para mostrar que es la respuesta correcta.

(PAUSE THE DVD. Make sure the class knows what to do and how to mark their answers. For the rest of this section, you will guide the students in English.)

(Record Start time: _____)

(Pace the children through items 1 – 20 using the directions below. This should take about 20 minutes. Allow enough time for students to answer each question.

You do not have to use the exact wording below, but please keep the class working together. Pacing is important to keep students from rushing through the items or spending too much time on one item.)

Turn to the next page. You should be on the page with the **kites** across the top.

P3. Put your marker under the **flag**. Fill in the circle under the picture that best completes the puzzle.

P4. Put your marker under the **lamp**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **pigs** across the top.

1. Put your marker under the **house**. Fill in the circle under the picture that best completes the puzzle.

2. Put your marker under the **fish**. Fill in the circle under the picture that best completes the puzzle.

3. Put your marker under the **elephant**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **whales** across the top.

4. Put your marker under the **crown**. Fill in the circle under the picture that best completes the puzzle.

5. Put your marker under the **flag**. Fill in the circle under the picture that best completes the puzzle.

6. Put your marker under the **lamp**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **butterflies** across the top.

7. Put your marker under the **house**. Fill in the circle under the picture that best completes the puzzle.

8. Put your marker under the **fish**. Fill in the circle under the picture that best completes the puzzle.

9. Put your marker under the **elephant**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **feathers** across the top.

10. Put your marker under the **crown**. Fill in the circle under the picture that best completes the puzzle.
11. Put your marker under the **flag**. Fill in the circle under the picture that best completes the puzzle.
12. Put your marker under the **lamp**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **pigs** across the top.

13. Put your marker under the **house**. Fill in the circle under the picture that best completes the puzzle.
14. Put your marker under the **fish**. Fill in the circle under the picture that best completes the puzzle.
15. Put your marker under the **elephant**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **whales** across the top.

16. Put your marker under the **crown**. Fill in the circle under the picture that best completes the puzzle.
17. Put your marker under the **flag**. Fill in the circle under the picture that best completes the puzzle.
18. Put your marker under the **lamp**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **butterflies** across the top.

19. Put your marker under the **house**. Fill in the circle under the picture that best completes the puzzle.
20. Put your marker under the **fish**. Fill in the circle under the picture that best completes the puzzle.

We have now completed the task. I will collect your test booklets now.

(Record end time: _____)

APPENDIX B

DIRECTIONS FOR ADMINISTERING VERBAL-DYNAMIC DIRECTIONS

Research Study Directions for Administration

**Cognitive Abilities Test
Practice Form
Grades 1-2**

Verbal Dynamic Directions

Introduction

The *Cognitive Abilities Test*[™] (*CogAT*[®]) is an integrated series of tests that appraises the cognitive development of students in kindergarten through grade 12. It assesses learned reasoning abilities in three areas of cognitive processing that are essential for success in school—verbal, quantitative, and nonverbal (or figural) reasoning.

Riverside Publishing is conducting a research study in first and second grade to try out this practice test which uses new types of test directions for *CogAT*. This research study will help us develop new forms of test directions and practice for *CogAT* that will help all students, but especially English-language learners, to perform their best on the test. We appreciate your participation in this study.

Please study this manual in advance of the testing day so that there will be no hesitation on your part when the test is administered.

For these “Verbal Dynamic” directions, it is important that you take time during the DVD instructions (the script indicates when to pause the DVD) to help students think through the problem and understand the matrix analogy format. You may need to branch off from the script depending on what the students say or whether they understand what to do. The important thing is that you encourage the students to put the rule for each puzzle into words and to describe how that gets them to the answer. For example, on sample 2 the rule is “the circle gets smaller, so the square must get smaller. That means the answer is D, a smaller square”. Students may need a minute or two of talking about the task before they understand this strategy.

If you speak Spanish fluently, you may go through this portion in both languages as necessary and permit answers in Spanish.

You will be administering one practice test which will take about 30 minutes, but you may want to allow more time, in case it is needed.

Preparing for Testing

- There should be one copy of this *Directions* manual for you and one test booklet for each student. Please note that different classrooms will receive different forms of the test. If you need additional copies of the student booklets, make sure that any copies you borrow have the exact same cover as the rest in your class. Make sure that your *Directions* manual is appropriate for the student booklets (i.e., the “Standard” Directions must be used in a classroom with “Standard or Cartoon” booklets. Store all of the test materials in a secure place.
- Be sure that each child has a soft, black-lead (No. 2) pencil with a good eraser. Have some extra pencils on hand in case they are needed.
- **Prepare a place marker for each child.** Place markers may be rectangular pieces of cardboard or folded pieces of paper, about 2 inches by 4 inches. **Place markers are not included in the test materials and must be prepared by the test administrator or an aide prior to testing.**
- **The test directions require playing a DVD for the class.** Arrange to have a TV with a DVD player on the day of testing. **Watch the DVD ahead of time while following along in the test**

booklet so that you know what special requirements your set of directions have.

Preparing student test booklets:

Student information is required on the cover of the test booklet. This information should be completed by the teacher. This information may be completed before or after testing, but must be complete before you return the materials to the testing coordinator.

- Student names are not being collected on the test booklets to protect student privacy. However, you may write the **first name of the student** on the cover so that they complete the right booklet.
- Instead of student names, record each student's **state ID number** on the front of the test booklet.
- Record the student's ELL status using the most recent English proficiency results.

During Testing

Follow the directions for administering the subtests **exactly** as they appear in this manual. You are provided with a full script of the DVD in green and orange print. Read aloud only the test directions printed in black. Directions in black within parentheses are for the test administrator and are not to be read to students. Follow along in the script throughout the directions as there are times when you may need to pause the video to give students a chance to mark answers or respond to questions.

Make sure that students understand the test directions and the sample questions before you begin each test. Answer any questions and help any students having difficulty. **Do not start the test until students understand what they are supposed to do.** Once the testing begins, you may not answer questions about specific items. However, questions about procedures and the mechanics of test taking may be answered at any time.

Move around the room while students are taking each test to make sure that they are working on the correct subtest and are filling in the answer bubbles appropriately.

There are no time limits for the subtests you will administer. Pace the students as they answer each test question. Allow enough time for all students to attempt to answer each question before telling students to go on to the next question. **Carefully record the exact time you begin and end each subtest.** Space to record this information is provided in this manual within the directions for each subtest.

In the space provided, please report any incident that could have an adverse effect on the test results.

After Testing

- Check that all students completed all items.
- Inspect each answer document and erase any stray marks.
- Complete “Teacher Information – Grade 1-2 Short Form” document as fully as possible
 - Look at the DVD or its envelope to find the **Version of DVD used** in your classroom (either Standard, Cartoon, Verbal Dynamic, or Nonverbal Dynamic).
- Stack the booklets with the “Teacher Information... Form” on top and bind with the rubber band provided.
- Return all test materials (including DVD) to the building test coordinator or other designated person.

Part 2: Administering the Test

Figure Matrices – Verbal Dynamic Directions

Today we will solve some puzzles. Some puzzles will be easy to solve. Others will be difficult to solve, but you should try your best. Let's try some together. *(start DVD)*

Open your book. You should be on the page with the **keys** across the top.

Abre tu libro. Debes estar en la página con las **llaves** que atraviesan la parte de arriba.

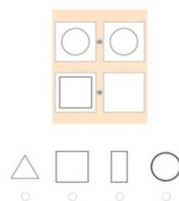
Put your marker under the **monkey**.

Pon tu cursor debajo del **chango**.

How would you complete this puzzle? Draw a shape in the empty box to finish the puzzle.

¿Como completarías este rompecabezas? Dibuja una figura en el cuadro vacío que complete el rompecabezas

(Encourage students to draw the answer into the empty box. You will hear a beep just before the silence ends. If students need more time to work, pause the video then.)



Now move your marker under the **broom**.

Ahora mueve tu cursor debajo de la **escoba**.

Does one of the pictures here match your answer?

¿Alguno de los dibujos es igual a tu respuesta?

Watch the video. Mira el video.

The shape has to be the same in each row.

La forma tiene que ser igual en cada hilera.

The square is the right answer. El cuadro es la respuesta correcta.

Look at your book and fill in the bubble under the square.

Mira tu libro y rellena el círculo debajo del cuadro.

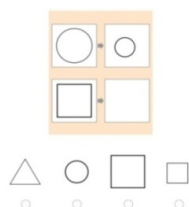
(pause DVD, if necessary, make sure students know to mark answer in bubble, restart DVD)

Now let's try another one. Put your marker under the **dog**.

Ahora intentemos otra. Pon tu cursor debajo del **perro**.

How would you complete this puzzle? Draw a shape in the empty box to finish the puzzle.

¿Cómo completaría este rompecabezas? Dibuja una figura en el cuadro vacío que complete el rompecabezas.



(At the beep, PAUSE DVD. Give students a moment to work THEN SAY:)

What was the rule for this puzzle? *(Hold up booklet and point to the two circles.)*

How did this circle become that circle? *(Encourage answers like 'it got smaller'.)*

How would you change the square in the same way? *(Encourage answers.)*

(point to answer choices) What would the answer look like?

Keep looking at your booklet. *(start DVD)*

Now move your marker under the **fish**.

Pon tu cursor debajo del **pescado**.

Does one of the pictures here match your answer?

¿Alguno de los dibujos es igual a tu respuesta?

Now watch the video.

Ahora mira el video.

The second circle is smaller than the first. What goes in the empty box?

El segundo círculo es más pequeño que el primero. ¿Qué va en el cuadro vacío?

a) Is this the answer? *¿Es esta la respuesta?* (encourage students to say yes/no)

No, the answer needs to be a square. *La respuesta tiene que ser un cuadro.*

b) Is this it? *¿Y esta?*

No, it is smaller, but it is not a square. *Es más pequeño, pero no es un cuadro.*

c) Is this the answer? ¿Es esta la respuesta?

No, it is a square, but it is not smaller than the one in the puzzle.

Es un cuadro, pero no es más chico que el cuadro que está en el rompecabezas.

d) Is this it? ¿Y esta?

Yes, because it is a square that is smaller than the one in the puzzle.

Sí. Es un cuadro más pequeño que el cuadro que está en el rompecabezas.

Look at your book and mark the right answer.

Mira tu libro y marca la respuesta correcta.

(pause DVD, make sure students know to mark answer in bubble, restart DVD)

Turn to the next page. There should be **kites** across the top.

Pasa a la página siguiente. Debería haber **papalotes** a través de la parte de arriba.

Put your marker under the **flag**.

Pon tu cursor debajo de la **bandera**.

How would you finish this puzzle? This time, don't draw your answer, just imagine it...

¿Cómo completarías este rompecabezas? Esta vez, no dibujes tu respuesta, simplemente imagínatela...

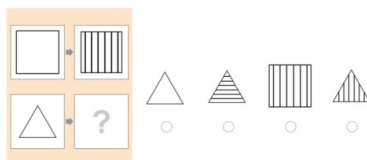
(At the beep, PAUSE DVD. Given students a moment to think THEN SAY:)

What is the rule for this puzzle? (Hold up booklet and point to the two squares left to right)

How did the square change from here to here? *(Encourage answers like 'added lines going up and down')*

Can we change the triangle the same way? What would that look like?

Keep looking at your book. *(start DVD)*



Now look at the row of pictures on the right *(point to options if there is confusion)*..

Ahora mira las hileras de dibujos en la derecha.

Which one shows your answer? Fill in the bubble.

¿Cual muestra tus respuestas? Llena el círculo.

Now watch the video. Ahora mira el video.

- a. Is this the answer? ¿Es esta la respuesta? (encourage students to say yes/no)

No, this is a triangle, but it doesn't have lines.

No. Este es un triángulo, pero no tiene las líneas.

- b. Is this it? ¿Y esta?

No, it is a triangle, but the answer needs to have lines going up and down

No. Es un triángulo, pero la respuesta tiene que tener líneas que van de arriba para abajo

- c. Is this the answer? ¿Es esta la respuesta?

No, the answer needs to be a triangle.

No, la respuesta tiene que ser un triángulo.

- d. Is this it? ¿Y esta?

Yes, because it is a triangle with lines going up and down.

Si porque es un triángulo con líneas que van de arriba para abajo.

Make sure you marked the right answer in your book.

Asegúrate de que hayas elegido la respuesta correcta en tu libro.

(Pause DVD, make sure students mark answer in bubble, restart DVD)

Let's try one more together.

Intentemos uno más juntos.

Put your marker under the **lamp**.

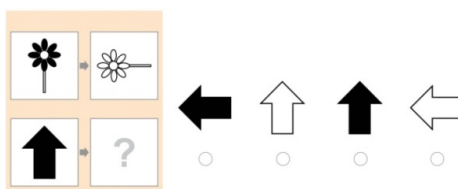
Pon tu cursor debajo de la **lámpara**.

How would you finish this puzzle? Imagine your answer.

¿Cómo completarías este rompecabezas? Imagina tu respuesta.

Remember, don't draw your answer anymore.

Recuerda, no dibujes más tu respuesta.



(At the beep, PAUSE DVD. Given students a moment to think THEN SAY:)

How did the flower change? *(Encourage students to come up with two rules like ‘it turned sideways’ and ‘it went from black to white’).*

(If necessary, prod:) Did the color change? What else happened?

That’s right, so the answer needs to be an arrow that is sideways and white.
(emphasize the three important features)

Keep looking at your book. *(start DVD)*

There are two rules—the flower was turned sideways but it also went from black to white.”

Hay dos reglas—la flor se volteó de lado pero también cambió de color de negro a blanco.

Now look at the row of pictures on the right.

Ahora mira la hilera de dibujos en la derecha

Which of the options shows an arrow turned sideways? Circle the arrows turned sideways.

¿Cuál de las opciones muestra una flecha volteada de lado? Circula las flechas volteadas de lado.

Look at the video if you are not sure. Mira el video si no estás seguro(a).

Which of the arrows is white? Circle the white arrows.

¿Cuál de las flechas es blanca? Circula las flechas blancas.

Only one arrow is BOTH turned sideways and white. So the last arrow is the answer.

Sólo una de las flechas esta volteada de lado y es de color blanco. Así que la última flecha es la respuesta.

Look at the video. The correct answer is marked.

Mira el video. La respuesta correcta está marcada.

If you marked the wrong answer in your book, you can change it now.

Si marcaste la respuesta incorrecta en tu libro, la puedes cambiar ahora.

(PAUSE THE DVD. Make sure the class knows what to do and how to mark their answers. For the rest of this section, you will guide the students in English.)

(Record Start time: _____)

(Pace the children through items 1 – 20 using the directions below. This should take about 20 minutes. Allow enough time for students to answer each question.

You do not have to use the exact wording below, but please keep the class working together. Pacing is important to keep students from rushing through the items or spending too much time on one item.)

Turn to the next page. You should be on the page with the **kites** across the top.

P3. Put your marker under the **flag**. Fill in the circle under the picture that best completes the puzzle.

P4. Put your marker under the **lamp**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **pigs** across the top.

1. Put your marker under the **house**. Fill in the circle under the picture that best completes the puzzle.

2. Put your marker under the **fish**. Fill in the circle under the picture that best completes the puzzle.

3. Put your marker under the **elephant**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **whales** across the top.

4. Put your marker under the **crown**. Fill in the circle under the picture that best completes the puzzle.

5. Put your marker under the **flag**. Fill in the circle under the picture that best completes the puzzle.

6. Put your marker under the **lamp**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **butterflies** across the top.

7. Put your marker under the **house**. Fill in the circle under the picture that best completes the puzzle.

8. Put your marker under the **fish**. Fill in the circle under the picture that best completes the puzzle.

9. Put your marker under the **elephant**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **feathers** across the top.

10. Put your marker under the **crow**. Fill in the circle under the picture that best completes the puzzle.

11. Put your marker under the **flag**. Fill in the circle under the picture that best completes the puzzle.

12. Put your marker under the **lamp**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **pigs** across the top.

13. Put your marker under the **house**. Fill in the circle under the picture that best completes the puzzle.

14. Put your marker under the **fish**. Fill in the circle under the picture that best completes the puzzle.

15. Put your marker under the **elephant**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **whales** across the top.

16. Put your marker under the **crow**. Fill in the circle under the picture that best completes the puzzle.

17. Put your marker under the **flag**. Fill in the circle under the picture that best completes the puzzle.

18. Put your marker under the **lamp**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **butterflies** across the top.

19. Put your marker under the **house**. Fill in the circle under the picture that best completes the puzzle.

20. Put your marker under the **fish**. Fill in the circle under the picture that best completes the puzzle.

We have now completed the task. I will collect your test booklets now.

(Record end time: _____)

APPENDIX C

DIRECTIONS FOR ADMINISTERING NONVERBAL-DYNAMIC DIRECTIONS

Research Study Directions for Administration

Cognitive Abilities Test Practice Form Grades 1-2

Nonverbal Dynamic Directions

Introduction

The *Cognitive Abilities Test™ (CogAT®)* is an integrated series of tests that appraises the cognitive development of students in kindergarten through grade 12. It assesses learned reasoning abilities in three areas of cognitive processing that are essential for success in school—verbal, quantitative, and nonverbal (or figural) reasoning.

Riverside Publishing is conducting a research study in first and second grade to try out this practice test which uses new types of test directions for *CogAT*. This research study will help us develop new forms of test directions and practice for *CogAT* that will help all students, but especially English-language learners, to perform their best on the test. We appreciate your participation in this study.

Please study this manual in advance of the testing day so that there will be no hesitation on your part when the test is administered.

Following the administration procedures in this manual exactly will ease the task of administering the test and will help ensure that students do not receive too much or too little assistance from the test administrator.

You will be administering one practice test which will take about 30 minutes, but you may want to allow more time, in case it is needed.

Preparing for Testing

- There should be one copy of this *Directions* manual for you and one test booklet for each student. Please note that different classrooms will receive different forms of the test. If you need additional copies of the student booklets, make sure that any copies you borrow have the exact same cover as the rest in your class. Make sure that your *Directions* manual is appropriate for the student booklets (i.e., the “Standard” Directions must be used in a classroom with “Standard or Cartoon” booklets. Store all of the test materials in a secure place.
- Be sure that each child has a soft, black-lead (No. 2) pencil with a good eraser. Have some extra pencils on hand in case they are needed.
- **Prepare a place marker for each child.** Place markers may be rectangular pieces of cardboard or folded pieces of paper, about 2 inches by 4 inches. **Place markers are not included in the test materials and must be prepared by the test administrator or an aide prior to testing.**
- **The test directions require playing a DVD for the class.** Arrange to have a TV with a DVD player on the day of testing. **Watch the DVD ahead of time while following along in the test booklet so that you know what special requirements your set of directions have.**

Preparing student test booklets:

Student information is required on the cover of the test booklet. This information should be completed by the teacher. This information may be completed before or after testing, but must be complete before you return the materials to the testing coordinator.

- Student names are not being collected on the test booklets to protect student privacy. However, you may write the **first name of the student** on the cover so that they

complete the right booklet.

- Instead of student names, record each student's **state ID number** on the front of the test booklet.
- Record the student's ELL status using the most recent English proficiency results.

During Testing

Follow the directions for administering the subtests **exactly** as they appear in this manual. You are provided with a full script of the DVD in green and orange print. Read aloud only the test directions printed in black. Directions in black within parentheses are for the test administrator and are not to be read to students. Follow along in the script throughout the directions as there are times when you may need to pause the video to give students a chance to mark answers or respond to questions.

Make sure that students understand the test directions and the sample questions before you begin each test. Answer any questions and help any students having difficulty. **Do not start the test until students understand what they are supposed to do.** Once the testing begins, you may not answer questions about specific items. However, questions about procedures and the mechanics of test taking may be answered at any time.

Move around the room while students are taking each test to make sure that they are working on the correct subtest and are filling in the answer bubbles appropriately.

There are no time limits for the subtests you will administer. Pace the students as they answer each test question. Allow enough time for all students to attempt to answer each question before telling students to go on to the next question. **Carefully record the exact time you begin and end each subtest.** Space to record this information is provided in this manual within the directions for each subtest.

In the space provided, please report any incident that could have an adverse effect on the test results.

After Testing

- Check that all students completed all items.
- Inspect each answer document and erase any stray marks.
- Complete "Teacher Information – Grade 1-2 Short Form" document as fully as possible
 - Look at the DVD or its envelope to find the **Version of DVD used** in your classroom (either Standard, Cartoon, Verbal Dynamic, or Nonverbal Dynamic).
- Stack the booklets with the "Teacher Information... Form" on top and bind with the rubber band provided.
- Return all test materials (including DVD) to the building test coordinator or other designated person.

Part 2: Administering the Test

Figure Matrices – Nonverbal Dynamic Directions

Today we will solve some puzzles. Some puzzles will be easy to solve. Others will be difficult to solve, but you should try your best. Let's try some together. *(start DVD)*

Open your book. You should be on the page with the **keys** across the top.

Abre tu libro. Debes de estar en la página con las **llaves** que atraviesan la parte de arriba.

Put your marker under the **broom**.

Pon tu cursor debajo de la **escoba**.

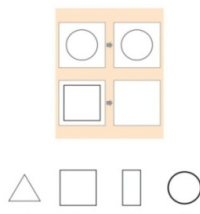
Now watch the video. *Ahora mira el video.*

This puzzle has a piece missing.

Le hace falta una pieza a este rompecabezas.

One of these pieces goes on the empty box to finish the puzzle.

Una de estas piezas va sobre el cuadro vacío para completar el rompecabezas.



Which one is the answer? *¿Cuál es la respuesta?*

(Long pause. There will be a beep right before the audio starts again. Pause DVD here if students need more time to think about the problem.)

Now watch the video. *Ahora mira el video.*

This is the correct answer.

Esta es la respuesta correcta.

Now look at your book. *Ahora mira tu libro.*

Fill in the bubble under the square to show it is the right answer.

Rellena el círculo debajo del cuadro para mostrar que es la respuesta correcta.

(pause DVD, make sure students know to mark answer in bubble; restart DVD)

Now let's try another one. Put your marker under the **fish**.

Ahora hay que intentar otro. Pon tu cursor debajo del **pescado**.

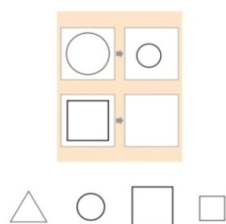
How would you complete this puzzle?

¿Como completarías este rompecabezas?

(Long pause. There will be a beep right before the audio starts again. Pause DVD here if students need more time to think about the problem.)

Now watch the video.

Ahora mira el video.



One of these pieces goes in the empty box to finish the puzzle.

Dibuja una figura en el cuadro vacío que complete el rompecabezas.

Which one is the answer?

¿Cuál es la respuesta?

This is the correct answer.

Esta es la respuesta correcta.

Look at your book and fill in the circle under the small square.

Mira tu libro y rellena el círculo debajo del cuadrado chico.

(pause DVD, make sure students know to mark answer in bubble)

Turn to the next page. There should be **kites** across the top.

Pasa a la siguiente pagina. Deben de haber **papalotes** atravesando la parte de arriba.

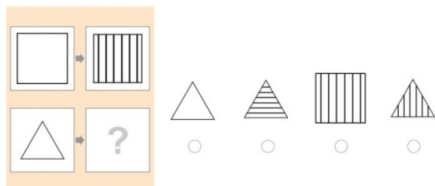
Put your marker under the **flag**.

Pon tu cursor debajo de la **bandera**.

How would you complete this puzzle?

¿Como completarías este rompecabezas?

(Long pause. There will be a beep right before the audio starts again. Pause DVD here if students need more time to think about the problem.)



Now watch the video.

Ahora mira el video.

Which one is the answer?

¿Cuál es la respuesta?

This is the correct answer.

Esta es la respuesta correcta.

Look at your book and fill in the circle under the answer.

Mira tu libro y rellena el círculo debajo de la respuesta. (pause DVD, make sure students know to mark answer in bubble)

Let's try one more together.

Intentemos uno más juntos.

Put your marker under the **lamp**.

Pon tu cursor debajo de la **lámpara**.

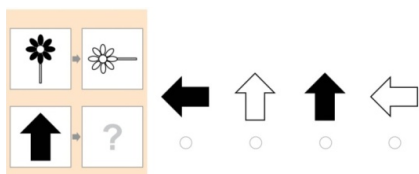
How would you complete this puzzle?

¿Como completarías este rompecabezas?

(Long pause. There will be a beep right before the audio starts again. Pause DVD here if students need more time to think about the problem.)

Now watch the video.

Ahora mira el video.



This is the correct answer. Esta es la respuesta correcta.

Look at your book and fill in the circle under the answer.

Mira tu libro y rellena el círculo debajo de la respuesta. (pause DVD, make sure students know to mark answer in bubble)

(PAUSE THE DVD. Make sure the class knows what to do and how to mark their answers. For the rest of this section, you will guide the students in English.)

(Record Start time: _____)

(Pace the children through items 1 – 20 using the directions below. This should take about 20 minutes. Allow enough time for students to answer each question.

You do not have to use the exact wording below, but please keep the class working together. Pacing is important to keep students from rushing through the items or spending too much time on one item.)

Turn to the next page. You should be on the page with the **kites** across the top.

P3. Put your marker under the **flag**. Fill in the circle under the picture that best completes the puzzle.

P4. Put your marker under the **lamp**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **pigs** across the top.

1. Put your marker under the **house**. Fill in the circle under the picture that best completes the puzzle.

2. Put your marker under the **fish**. Fill in the circle under the picture that best completes the puzzle.

3. Put your marker under the **elephant**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **whales** across the top.

4. Put your marker under the **crow**. Fill in the circle under the picture that best completes the puzzle.

5. Put your marker under the **flag**. Fill in the circle under the picture that best completes the puzzle.

6. Put your marker under the **lamp**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **butterflies** across the top.

7. Put your marker under the **house**. Fill in the circle under the picture that best completes the puzzle.

8. Put your marker under the **fish**. Fill in the circle under the picture that best completes the puzzle.

9. Put your marker under the **elephant**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **feathers** across the top.

10. Put your marker under the **crown**. Fill in the circle under the picture that best completes the puzzle.
11. Put your marker under the **flag**. Fill in the circle under the picture that best completes the puzzle.
12. Put your marker under the **lamp**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **pigs** across the top.

13. Put your marker under the **house**. Fill in the circle under the picture that best completes the puzzle.
14. Put your marker under the **fish**. Fill in the circle under the picture that best completes the puzzle.
15. Put your marker under the **elephant**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **whales** across the top.

16. Put your marker under the **crown**. Fill in the circle under the picture that best completes the puzzle.
17. Put your marker under the **flag**. Fill in the circle under the picture that best completes the puzzle.
18. Put your marker under the **lamp**. Fill in the circle under the picture that best completes the puzzle.

Turn to the next page. You should be on the page with the **butterflies** across the top.

19. Put your marker under the **house**. Fill in the circle under the picture that best completes the puzzle.
20. Put your marker under the **fish**. Fill in the circle under the picture that best completes the puzzle.

We have now completed the task. I will collect your test booklets now.

(Record end time: _____)

APPENDIX D

STUDENT TEST BOOKLET FOR STANDARD DIRECTIONS

Teacher Name _____

Student ID # _____

Gender: Boy

Girl

ELL status (choose one): Non-ELL

FEP (CELA 5)

LEP (CELA 3 or 4)

NEP (CELA 1 or 2)

Cognitive Abilities Test™ (CogAT®)

Research Study Student Booklet Grade 1-2 Practice Form

*Use for
Standard or Cartoon
Directions Only*





Figure Matrices

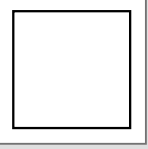
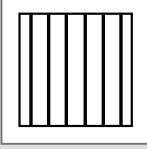
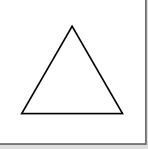

P1


P2


Figure Matrices

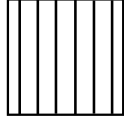
P3

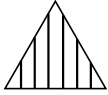
   	

	→	
	→	




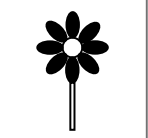
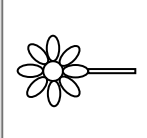
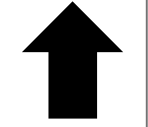



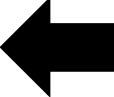


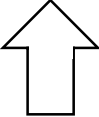



P4

	→	
	→	







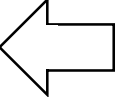
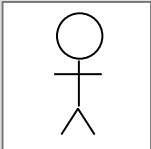


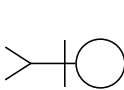
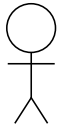


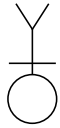
Figure Matrices


1


		→		
		→	?	





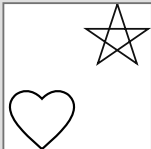





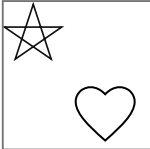





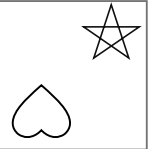
2


		→		
		→	?	



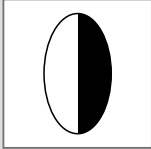
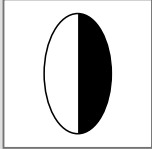
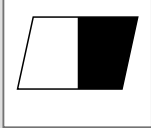












3

		→		
		→	?	











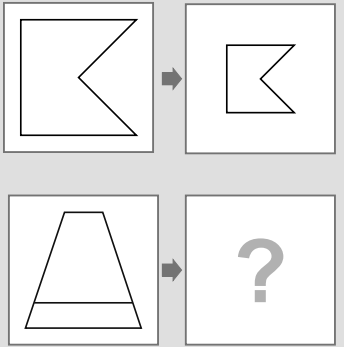
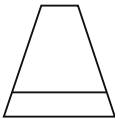
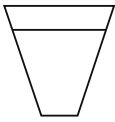



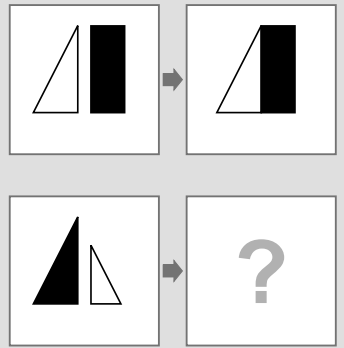






Figure Matrices

4

		 <input type="radio"/>	 <input type="radio"/>	 <input type="radio"/>	 <input type="radio"/>
---	---	--	---	--	--

5

		 <input type="radio"/>	 <input type="radio"/>	 <input type="radio"/>	 <input type="radio"/>
---	--	---	--	---	---

6


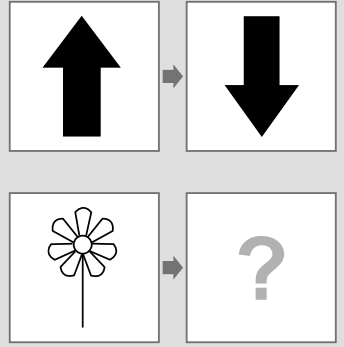
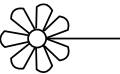

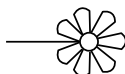
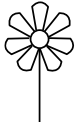
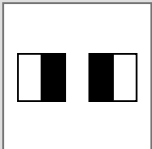
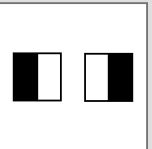
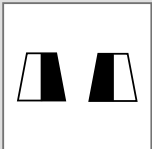
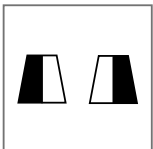
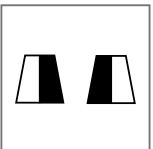
		 <input type="radio"/>	 <input type="radio"/>	 <input type="radio"/>	 <input type="radio"/>
---	---	--	---	--	--

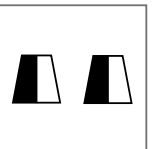

Figure Matrices

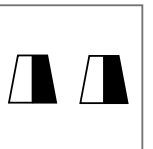
7

	→	
	→	?

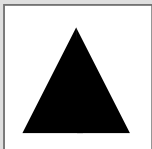

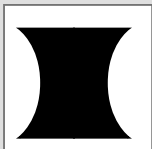

○



○

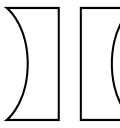

○



○


8

	→	
	→	?

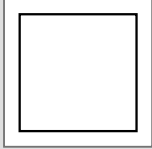
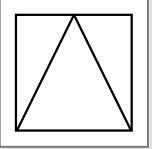
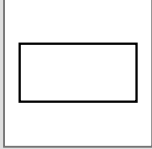

○

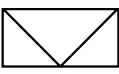

○

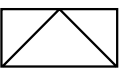

○

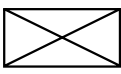

○

9

	→	
	→	?


○


○


○



○

Figure Matrices

10

		→			
		→	?		
				○	○
				○	○

11

		→			
		→	?		
				○	○
				○	○

12

		→			
		→	?		
				○	○
				○	○

Figure Matrices

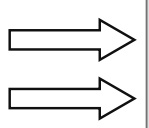

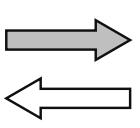
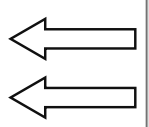


13



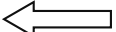

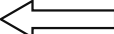
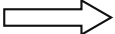

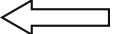
14

15


Figure Matrices

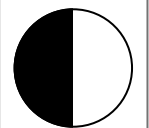

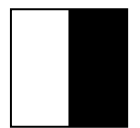



16





		
		

- | | | | |
|---|--|---|---|
|  |  |  |  |
|  |  |  |  |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |



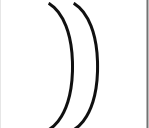

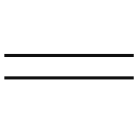



17

- | | | | |
|--|---|--|--|
|  |  |  |  |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |



18





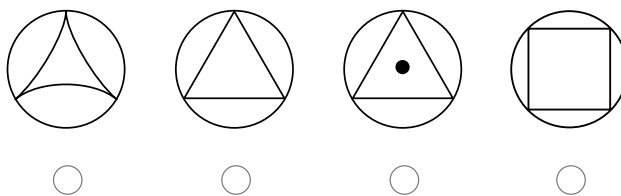
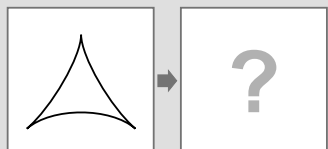
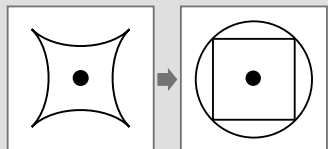
- | | | | |
|---|--|---|---|
|  |  |  |  |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

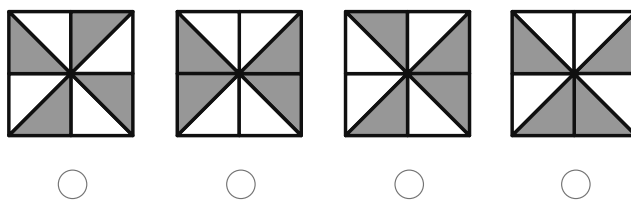
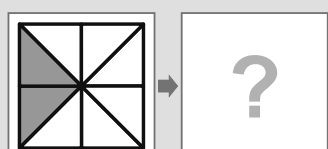
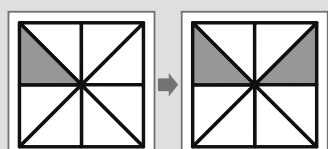



Figure Matrices

19



20



APPENDIX E

STUDENT TEST BOOKLET FOR VERBAL- OR NONVEBAL-DYNAMIC DIRECTIONS

Teacher Name _____

Student ID # _____

Gender: Boy

Girl

ELL status (choose one): Non-ELL

FEP (CELA 5)

LEP (CELA 3 or 4)

NEP (CELA 1 or 2)

Cognitive Abilities Test™
(CogAT®)
Research Study
Student Booklet
Grade 1-2 Practice Form

Use for
Verbal Dynamic or Nonverbal Dynamic
Directions Only

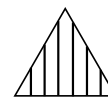
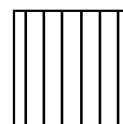
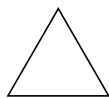
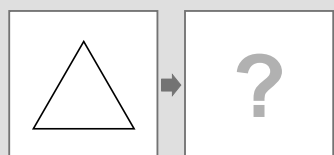
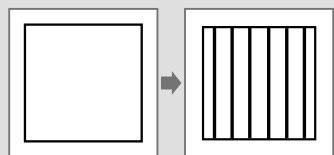
Figure Matrices

P1

P2

Figure Matrices

P3



P4

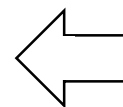
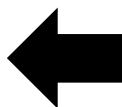
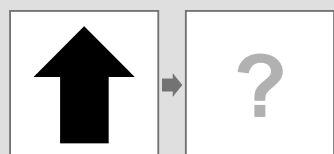
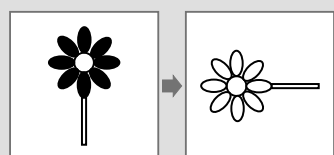
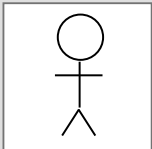

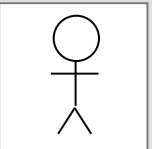
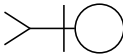
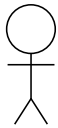
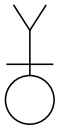

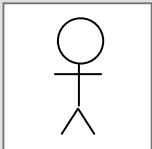





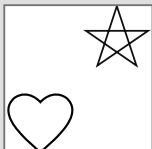









Figure Matrices


1

1	  				
	  	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

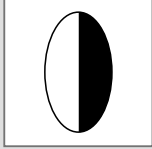

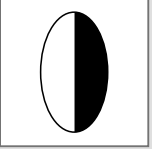









2

2	  				
	  	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



3

3	  				
	  	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



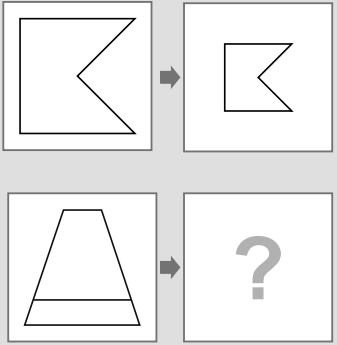
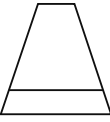
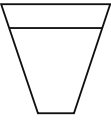



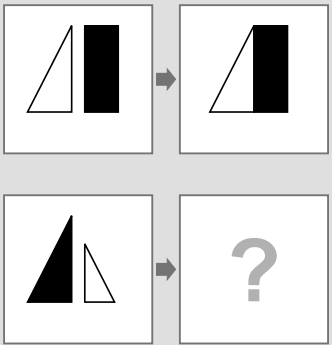






Figure Matrices

4

		 <input type="radio"/>	 <input type="radio"/>	 <input type="radio"/>	 <input type="radio"/>
---	---	--	---	--	--

5

		 <input type="radio"/>	 <input type="radio"/>	 <input type="radio"/>	 <input type="radio"/>
---	--	---	--	---	---

6


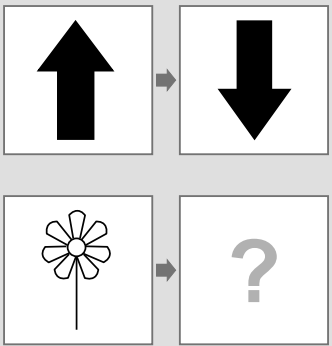
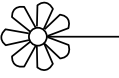

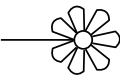
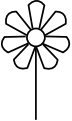
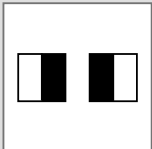
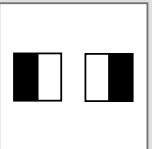
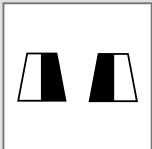
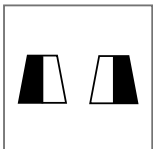
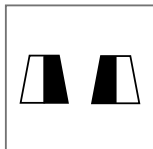
		 <input type="radio"/>	 <input type="radio"/>	 <input type="radio"/>	 <input type="radio"/>
---	---	--	---	--	--

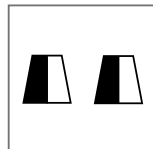

Figure Matrices

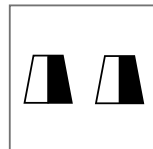
7

	→	
	→	?

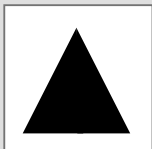

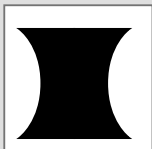

○

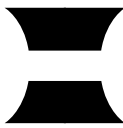

○

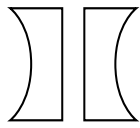

○



○


8

	→	
	→	?

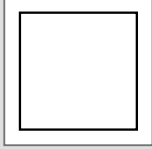
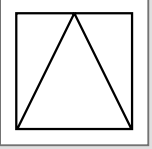
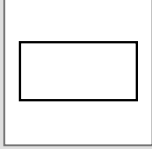

○



○



○



○

9

	→	
	→	?


○


○


○

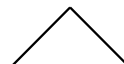

○

Figure Matrices

10

	→	
	→	

-
-
-
-



11

	→	
	→	

-
-
-
-



12

	→	
	→	

-
-
-
-



Figure Matrices

13

13

○ ○ ○ ○

14

14

○ ○ ○ ○

15

15

○ ○ ○ ○