INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

METHODS FOR ASSESSING STUDENT LEARNING IN THE STATE OF ARIZONA

by

Stephen Jay Midyett

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2001

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

THE UNIVERSITY OF ARIZONA ®
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by Stephen Jay Midyett

entitled Methods for Assessing Student Learning in the State of Arizona

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

Jerome D'Agostino	10/19/01 Date
Patricia Jones	10/19/2001 Date
Darrell Sabers	10-19-01 Date
	Date
	Date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copy of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Dissertation Director Darrell Sabers — 10-19-01 Date

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under the rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of sources is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarships. In all other instances, however, permission must be obtained from the author.

SIGNED: 

ACKNOWLEDGEMENTS

I wish to express my appreciation to Dr. Darrell Sabers, my mentor, for his guidance throughout the preparation of this dissertation as well as my entire graduate career. I would like to thank Dr. Pat Jones for her invaluable input into the procedures of this dissertation and for the countless hours I spent in her statistics courses. I would also like to thank Dr. Jerry D'Agostino for his assistance and time during this process.

I would like to express my thanks to my wife, Lisa, for her patience and encouragement during this long process. She took over many of my responsibilities allowing me to pursue this dissertation. I wish to thank all of my family members for their love and support, in particular my mother, Judith, who despite enduring the death of my father last year, remained positive and provided me with encouragement.

Finally, I would like to offer special thanks to my father who died during the writing of this paper. He taught me the meaning of hard work and persistence. He is, and will be, sorely missed.

TABLE OF CONTENTS

TABLE OF CONTENTS - *Continued*

TABLE OF CONTENTS - *Continued*

TABLE OF CONTENTS - *Continued*

LIST OF TABLES

LIST OF TABLES – *Continued*

LIST OF TABLES – *Continued*

LIST OF TABLES – *Continued*

LIST OF FIGURES

LIST OF FIGURES – *Continued*

LIST OF FIGURES – *Continued*

LIST OF EQUATIONS

ABSTRACT

The effectiveness of a method using scaled scores and a correction for regression to the mean (RTM) designed to measure academic growth attributable to schools was compared to several alternative methods all incorporating simple (unadjusted) growth. Problems with scaled scores and the correction for RTM were discussed.

Three alternative methods using normal curve equivalent (NCE), percentile rank (PR), and stanine scores were presented and compared to the scaled score method. A variation of the scaled score method without the correction for RTM was proposed to examine the effects of the correction. Two variations of the NCE and PR score methods were constructed with adjusted passing criteria to examine the effect of accounting for measurement error.

Matched-student (1998-1999) Stanford 9 Achievement Test scores from the State of Arizona were used to compute a dichotomous one year's growth indicator (OYG) and a five-point within-state rank-ordered growth indicator (the Star Rating) for each school/grade unit using each of the proposed methods.

Results showed that the methods using NCE or PR scores were more likely than the method using scaled scores to assign the same OYG decision to each school/grade unit. The correction for RTM resulted in school/grade units with low initial status having to (inappropriately) make more than one year's worth of growth to achieve a passing OYG decision. The results tended to confirm correlations between initial status and the

simple growth indicators in the alternative methods, but for a majority, the magnitudes of the correlations were not large enough to warrant dismissing simple growth.

Recommendations from the study were: 1) Scaled scores and the correction for RTM should not be used in any of the methods; 2) Methods that account for error should be used to allow for control over the possibility of misidentification of failing schools as well as the proportion of schools that are identified as needing assistance; 3) The current minimum unit size criterion of eight students should remain, because increasing the number would result in too many units not included in analyses.

CHAPTER 1

INTRODUCTION

This research project is focused on the methodology used to assess schools in terms of student growth. In particular, the research is focused on the process of developing a measure of academic growth in the State of Arizona. Technical issues with Arizona's methods are described. Alternative methods for assessing academic growth are offered, and the research leading to a recommended method is detailed.

Changing from Measures of Status to Growth

Like many other states, Arizona has used standardized tests to assess an individual student's achievement for many years. Arizona implemented the Stanford 9 achievement tests in 1997; prior to that, the Iowa Test of Basic Skills had been used for 12 years. The typical application of these scores was to provide an indicator of the standing of each student. For assessments at the school level, student scores were averaged for each grade level in the schools and used as an indicator of the status of a particular school/grade unit. Knowing the status of a school/grade unit can be useful. For example, the averaged scores for school/grade units could be ranked across the state, thus allowing the identification of units with exceptionally good or poor performance records.

Schools and school districts have always been accountable to some degree for demonstrating student progress. In the past in Arizona, the metric used for progress was an annual status indicator. Year-to-year comparisons of averaged school/grade level scores have served as a rough measure of annual academic growth. There are problems

with comparing the status from year-to-year as the growth estimates can be confounded by factors beyond school control. For example, the achievement scores from student transfers into a school just prior to the administration of the test will not accurately reflect the effectiveness of the school. Other factors that make the comparison of year-to-year status a questionable practice are described in the Chapter 2.

New mandates for school accountability from the State of Arizona have created even a greater need to identify the role schools played in the education of students. In November of 2000, voters in Arizona passed the controversial Proposition 301. Under the new statute, school districts are responsible for providing evidence that their students are maintaining adequate yearly progress. The evidence is to be provided in the context of a state accountability system to be developed by the Arizona Department of Education. A school not maintaining adequate progress will be labeled initially as an "Under-performing School." If the school is under-performing for a second year, it will be labeled a "Failing School." Districts with failing schools will face state board review, possibly receive reduced funding, and be required to publish their status in newspapers and on voting ballots.

A few years earlier, the Arizona Department of Education anticipated the emphasis on accountability and began the groundwork for a system to deal with it. In particular, they recognized that Arizona needed an appropriate measure of growth in addition to the status indicator already in use. In 1998 a new approach to assessing the growth of students in each school/grade unit was created. The first report to include this

new approach was completed in February of 2000 and was called the Arizona Measure of Academic Progress (MAP).

The primary goal of the MAP report was to assess accurately the effectiveness of each school/grade unit in educating students. To achieve the goal, methods were examined which would limit the effects of factors beyond the control of the school as much as possible. In addition, there was a heavy emphasis that resulting growth indicators were publicly consumable, that is, easy to understand, minimizing the risk of score misuse or misinterpretation. The process would not be a simple task.

Other Accountability Systems

The problems encountered in Arizona for measuring the impact of the school on education are not new. Many states currently continue to struggle with these issues as they have done in the past. Three approaches have been selected to demonstrate the variety in assessment programs, all of which were prompted by accountability mandates.

The Oregon Approach

In Oregon, teacher effectiveness is the primary focus of the state-mandated assessment system. The point is to let parents know the quality of the new teachers. Under the Oregon Teacher Work Sample Methodology, student teachers are required to complete a series of pre-defined steps to evaluate the needs of the students, to design and implement a curriculum, and to assess student learning as a result of instruction. Student teachers construct, administer, and score a pre-instructional assessment of the content to be taught, administer a 3 to 5 week unit of instruction, and then re-assess students to

obtain gain scores for each student. A licensure committee reviews the work by the student teacher. Acceptance of the work leads to certification of the teacher.

Multiple measurements are collected for the program. Student-based measures collected in the Oregon program include pre and post percentage correct scores from students on the teacher-made tests. Teacher-based measures include a "teacher-skills" score based on observation and review of required materials collected during the internship period.

The measures are aggregated in the Oregon program by computing simple gain scores for the students to assess academic growth. Pre to post gain is evidence of academic progress for the students. Teacher profile reports are generated from the competency-based teacher-skills scores. These profiles are made available for parents to review.

Advocates of the Oregon program (Schalock, Schalock, & Girod, 1997) highlight four major features of the program. First, the assessment program appears reasonable to teachers, parents, and school administration. Second, it is feasible to implement, meaning it is applicable to any new teacher in any teaching situation. The third feature is that it can serve multiple purposes, by allowing an evaluation of student growth and acting as a training vehicle for teachers. Finally, advocates believe evidence collected thus far demonstrates notable learning gains as well as reproducible findings.

Critics of the Oregon program (Airasian, 1997) point out all the major criticisms that arise when standardized assessments are not used. They highlight the lack of reliability, content validity issues, and general lack of quality of the teacher-made

assessments. These issues lead to the argument that using the assessments in this system, for high-stakes purposes, would prove to be indefensible.

The Dallas Approach

In the Dallas Value-Added Accountability System the main evaluation focus is on both teachers and school effectiveness. The main point is to let the public know how the teachers and schools are performing in terms of academic achievement from the students. Annual assessment profiles are generated for both teachers and schools. Teacher assessments are reported publicly only in aggregated form, while individual school results are public.

Student-based measures in the Dallas model include a national standardized achievement test, The Iowa Test of Basic Skills, and a state-mandated criterion-referenced test, The Texas Assessment of Academic Skills. There are no direct teacher-based (teaching competency) measures made in the program. Other student and school level variables such as gender, language proficiency, and socioeconomic status are collected for use as covariates in statistical analyses.

Student learning in the Dallas model is based on the statistical weighting of the national norm-referenced test and the state-based criterion-referenced test. In particular, two Hierarchical Linear Model (HLM) analyses are performed, one model for teachers and another for schools. In the two-stage HLM Teacher model, the first stage is to compute predicted growth for students using student level covariates such as gender and ethnicity. The second stage of the HLM model is to adjust the predicted growth estimates from the first stage by teacher level covariates such as classroom mobility and average

family education. Teacher effectiveness is based on the degree to which the teacher exceeds the predicted growth. In the two-stage HLM School model, the first stage is the same as it is for the teacher model. The second stage in the school model is to adjust the predicted growth estimates from the first stage by school level covariates such as school mobility and school-level free or reduced lunch. School effectiveness is based on the degree to which the school exceeds the predicted growth. Webster (2000) provides a complete listing of the covariates in the models.

Advocates of the Dallas program (Thum & Bryk, 1997) point out the advantages of controlling covariates that may otherwise confound the estimates of student learning. They maintain that certain student populations are more difficult to educate than others. If teachers and schools are going to be held accountable, then a level playing field needs to be created.

Critics of the Dallas program (Sykes, 1997) argue that a testing system that emphasizes discrete items in multiple-choice format and covering only basic skills leaves out attention to complex cognitive performances and tasks. Such testing leads to a narrowing of the subject content taught and thus learned.

The Tennessee Approach

In the Tennessee Value Added Assessment System the main evaluation focus is on several levels ranging from student-level to the district level. The program is designed to evaluate effectiveness of students, teachers, school/grade levels, schools, and districts in terms of student academic achievement. Annual assessments are conducted for all

levels of analysis. Student and teacher-level reports are confidential, but school and district-level reports are publicly available.

The student-based measures in the Tennessee model are the scaled scores from a state norm-referenced exam, the Tennessee Comprehensive Assessment Program. Like the Dallas model there are no direct teacher-competency measures made in the program. Other student and school level variables such as gender and socioeconomic status are collected for use as covariates in analyses.

The Tennessee model estimates student learning and school effect through a mixed-model statistical procedure that attempts to control for as many sources of variability that can be measured in the system. Any variables linear (equal-interval) in their metrics, highly correlated with curricular objectives, and possessing appropriate measurement sensitivities can be used in the model. Generally speaking, the effect on student learning for all of the covariates in the model is estimated as well as the effects from particular teachers or schools or districts.

Advocates of the Tennessee program (Sanders, Saxton, & Horn, 1997) argue that controlling for as many factors as possible is really the only way to produce fair or unbiased estimates of learning. Desirable features of the model include the capability to include multiple years or sources of data from a single student to produce academic growth estimates. In addition, the mixed-model can handle missing data-points that might cause problem in linear-regression based models such as HLM.

Critics of the Tennessee program (Walberg & Paik, 1997) cite disadvantages of the program. First, the statistical methods are complex and not sufficiently detailed for

replication. The complexity can make the interpretation of results questionable. Second, it is argued that the costs involved in making so many estimates are not proportional to the benefits received. Finally, they point out that there are few mechanisms in the program which link measured effects with instructional change.

The Arizona Measure of Academic Progress

In developing the Arizona Measure of Academic Progress (MAP), the Arizona Department of Education (ADE) initially considered the model used in Tennessee. The developers of the Tennessee program offered an analysis service. However, cost proved to be too high to fit within the Arizona budget. In addition to costs, the data required for the statistical model were not currently being collected on a statewide basis. Adjusted growth models such as the Dallas HLM model were not considered for two reasons. First, the additional data required were not collected at the state-level. Secondly, developers at the ADE believed that the more complex adjusted growth models would not be used by the teachers and school administration. They feared that teachers would see the process as a "black box" and not feel comfortable using or trusting the data. As a result, the ADE decided to focus on models of simple growth.

The first MAP report was generated in 2000. The general approach taken was to estimate student learning from scaled scores from the Stanford 9 Mathematics and Reading achievement tests. Then, national norms were used to determine whether grade levels within schools had achieved one year's worth of growth. Schools were then ranked based on their relative performances.

The initial MAP program had little in common with any of the other models mentioned (Oregon, Dallas, or Tennessee). Revisions planned for the MAP 2001 and later reports may include a weighted combination of the national achievement test and the "AIMS", a state criterion-referenced test. While these planned changes may lead to a program that more closely resembles the Dallas system, which also includes a weighted combination of a norm-referenced and a criterion-referenced test, that system is not specifically targeted as a model.

The larger context of how the Arizona MAP system will evolve is certainly important but is beyond the scope of this research project. This research focuses on the specific task of producing the best growth indicator from the standardized achievement test given in Arizona. The impetus for this research was the methodology used in the initial report.

The initial MAP report estimated the academic growth made by Arizona schools from 1998 to 1999. Stanford 9 achievement data collected on students from these years were used as the basis for assessing academic progress. Not all grade levels were included in the analysis. Problems with curricular fit between the Stanford 9 tests and the content in grades 9 to 12 led to the omission of these grades, and only grades 3 to 8 were included. The following is an overview of the methodology used in that report.

Student scores were included in the analyses if four criteria were met: 1) the student took the exam in the same school in both 1998 and 1999; 2) no special testing accommodations were provided to the student; 3) the student did not re-take the same

level exam the second year; 4) the student had a valid score in the subject area for both years.

The results of the analysis were reported for each grade level in a school. A school/grade level was omitted from analysis if it met any of the following conditions: 1) there were fewer than 8 students in the unit; 2) less than 25% of eligible students in a grade level were matched across consecutive years; 3) the grade level did not have scores on record for both years; 4) the school did not contain at least two adjoining grade levels between third and eighth grade.

Academic growth was computed by finding the difference between the average scaled score performance from one grade to the next. This growth was adjusted to compensate for the effects of Regression to the Mean (RTM). The adjusted growth was then used to produce two indicators. The first was One Year's Growth (OYG), which was computed by comparing the adjusted growth to expected growth amounts for students at the 50^{th} percentile. A school/grade unit which had an adjusted growth mean equal to or greater than the expected growth was said to achieve OYG. The second indicator was called the "Star Rating." For this, quintiles were assigned to each school based on adjusted growth in each grade level within Mathematics or within Reading. The first MAP report contained the yes/no decision of achieving OYG and the quintile (The Star Rating) for each grade level within a school.

Potential Problems with the MAP

There are three main issues with Arizona's initial method that make the results questionable. The first is the mismatch between the intended goal of the analysis and the

unit of analysis used in the report. The second is the use of scaled scores for making decisions about academic growth relative to the nation. The third is the correction for the effects of RTM.

One of the goals of Arizona's first MAP report was to examine the academic growth of individual students over time. In Arizona's initial method, growth is computed as a function of the means of each school/grade unit. The academic growth of an individual student is never computed. The unit of analysis is important because additional information can be gained by computing growth at the individual student level.

There are several issues related to the scaled scores that make their use suspect. The first point is that the Stanford 9 scaled scores are based on analysis using the Rasch model. Past research will be presented that demonstrates that the one-parameter IRT model does not fit multiple-choice test data well. Secondly, vertical equating carried out using the Rasch model may be questionable. Finally, the scaled scores were used in Arizona's method to make judgments about norm-referenced academic growth. Scaled scores require a norm-referenced expected growth value to assess whether a student maintains the same national percentile ranking from one year to the next. It is demonstrated in this research that scaled scores proved to be a problem in the Arizona method.

The state's method included an adjustment to correct for the effects of RTM. The correction is inappropriate for this particular analysis. It is argued that because of the manner of selection of students into each unit of analysis, the standard correction for RTM is not applicable. Inappropriately correcting for RTM can lead to dismissing real

educational effects as statistical artifacts. The correction systematically leads to reducing the amount of growth needed to maintain a high percentile rank. It also leads to increasing the amount of growth needed for those at a lower percentile rank.

Goals of this Research

This current research project had several objectives. The first was to examine Arizona's initial methodology in a critical but constructive manner. In doing so, potential problems with that method were detailed. The second objective was to propose alternative approaches for conducting the analysis of school effectiveness. By building on the sound portions of Arizona's method and incorporating existing standards, new approaches were offered. The third objective for the current research was to conduct an empirical comparison of the State's initial method and the proposed alternative methods. From this comparison, the differences between the methods were explained, and the potential effects the differences could have in practice were discussed.

The ultimate goal of this current research was the recommendation of one or more sound methods for assessing school effectiveness from student achievement. It is important for Arizona to have a sound method in place because it will serve as the basis for a more comprehensive evaluation of instructional effectiveness.

CHAPTER 2

RELATED RESEARCH

In this section, the methods used for the first Arizona MAP report are listed and examined in detail. Measurement concepts related to the process are presented with supporting research, as well as additional considerations relevant to measuring student learning. Three alternative methods for computing student learning are proposed. Finally, performance differences between Arizona's initial method and the three alternative methods are discussed. Research questions that were generated to examine differences between the methods are enumerated and described.

Part I: Arizona's Initial Method

Arizona's method is explained in four components or steps in order to facilitate the discussions regarding each step. The steps were completed separately for two subject areas (Reading and Mathematics). The four major components and the processes in each are as follows:

Step 1. Student scores were matched across two consecutive grade levels to ensure that each student in the analysis had two years of test scores for at least one subject area (Reading or Mathematics).

Step 2. Academic growth was estimated as follows: the scaled scores of students were averaged within each grade level (by subject area) within each school for both the first and second years. The pre-means (1998) were statistically adjusted to estimate the expected score on a second test. The difference between the post-means (1999) and the

estimated expected pre means for each school/grade unit was computed to estimate true growth. This difference was called the adjusted growth score for Reading and Mathematics.

Step 3. The OYG decision was made for each school/grade unit by comparing the adjusted growth amount to an expected growth amount. If the adjusted growth was larger than or equal to the expected amount, then the school/grade unit achieved OYG. If the adjusted growth was less than expected, the school/grade unit did not achieve OYG.

Step 4. A within-state ranking indicator called the Star Rating was computed. The adjusted growth values for all schools at the same grade levels were ranked and quintiles were produced. The quintile assigned to each school indicated the number of "Stars" in Reading or Mathematics for each school/grade unit.

Matched Samples and Inclusion Criteria

The first step in the Arizona method was a student record matching process that ensured that the students included in the MAP analysis had scores for the Stanford 9 in both 1998 and 1999. A combination of student identification numbers, last name, first name, date of birth, and gender, were used to match the pre-measure (1998) with the post-measure (1999). The matched student records were included in the analysis only for students who took both tests in the same school, did not require special accommodation in either year, did not re-take the same grade level exam in the second year, and had valid scores in the subject for both years.

Table 2.1 shows the matching rates and percent of eligible students rates for each grade level from the 1998-1999 MAP analysis. The columns labeled "Valid 1999 Scores"

are the number of students with valid Stanford 9 scores in either Reading or Mathematics in 1999. The columns labeled "Percent Matched with 1998 Scores" are the proportion of matched records from 1998 to 1999 out of the number of "Valid 1999 Scores". The columns "Percent of Valid 1999 Scores Used in Analyses" are the proportion of remaining students after the student inclusion and minimum units size criteria were implemented.

Table 2.1

Student Match Rate by Grade (1998 to 1999)

	Reading			Mathematics		
Grade Level 1999	Valid 1999 Scores	Percent Matched with 1998 Scores	Percent of Valid 1999 Scores Used in Analyses	Valid 1999 Scores	Percent Matched with 1998 Scores	Percent of Valid 1999 Scores Used in Analyses
4	59,514	80.22%	58.99%	61,560	79.49%	59.21%
5	60,980	80.51%	61.61%	62,351	79.71%	62.33%
6	59,197	83.78%	44.24%	60,153	83.15%	44.54%
7	59,190	84.35%	34.82%	59,542	84.15%	35.02%
8	57,207	85.32%	70.32%	57,338	85.20%	69.95%
Total	296,088	82.80%	53.94%	300,944	82.27%	54.18%

The values from Table 2.1 show a decrease in the percent of valid scores used in analyses for sixth and seventh grade. This decrease is due to one of the student inclusion criteria, (a student had to have taken the exam in the *same school* both years). This criterion is problematic for the transition into middle and junior high schools. Since the

original analyses, this criterion was modified to include students if they had been in the school for a full year and had a valid pre-measure from another school in Arizona.

The matching and selection process used in the Arizona analysis has some desirable features. Selection can be a threat to internal validity when an effect is due to the qualitative differences between groups (Cook & Campbell, 1979). The process of matching the students across both years ensures that the samples are essentially the same across years. While there may be some changes that occur in the students during a year, some factors such as gender and ethnicity will not change, and other factors such as socioeconomic status and school budgets are unlikely to change greatly within the course of a single year. From a research design point of view the matching makes the attribution of learning to the instruction more valid.

Including only those students who have been in the school for a full year ensures that a school/grade unit is credited with a student's improvement or deficit only if the school has had a reasonable amount of time to educate the students. Kerbow (1996) reported that the average achievement level in a school will improve once mobile students (students who change schools during a school year) are excluded from analyses.

The student matching process does not control for differences between the kinds of students found in one school/grade unit to another. For example, one school may serve a group of students who are different from another school in terms of gender or ethnicity. Other adaptations in the research design or statistical analysis may need to be taken if the results of one school are to be compared to another, such as the hierarchical multiple regression technique in the Dallas Value-Added Accountability System.

Matching is certainly a benefit in controlling the effects of selection within a school. However, while controlling selection aids in estimating the impact of the school in terms of instruction, the matching and selection criteria reduce the number of students included in an analysis, leading to an external validity problem. Results from the reduced sample may not generalize to the entire student body of a particular school. For the same school, there may be a difference in the growth estimated from a matched sample as opposed to the difference in two "status" measures for the same two years. For example, Table 2.1 shows that only 35% of the 1999 seventh graders with valid scores in either Reading or Mathematics were included in the 1998-1999 MAP analysis. These differences need to be made clear to teachers and school administrators who will be interpreting the results.

Growth in the Arizona Method

The second step of the Arizona analysis was to compute the simple growth for each school/grade unit. This was done in two parts. First, for each school year (1998 and 1999), the scaled scores from the Stanford 9 from every matched student in the unit were averaged. This process was done separately for Reading and Mathematics, yielding four separate scores. Second, the adjusted growth scores were calculated, controlling for the effects of RTM. The adjustment was made in the following manner. First, the 1998 scaled score means for every school/grade by subject area were standardized. The equations for the transformation are

$$z_{98sag} = (\bar{x}_{98sag} - \bar{\bar{x}}_{98.ag}) / \hat{\sigma}_{98.ag} \qquad (2.1)$$

where $\bar{x}_{98sag}$ = the mean scaled score for 1998 for school s in area a in grade g,

$\bar{\bar{x}}_{98.ag}$ = the mean scaled score for all schools in area a in grade g,

$$= \frac{1}{S_{ag}}\left(\sum_{s=1}^{Sag} \bar{x}_{98Sag}\right) \text{ where } S_{ag} \text{ is the number of schools in area } a \text{ in grade } g, \text{ and} \tag{2.2}$$

$\hat{\sigma}_{98.ag}$ = the standard deviation of the scaled score means for all schools in area a in grade g,

$$= \left(\frac{\sum_{s=1}^{Sag} (\bar{X}_{98sag} - \bar{\bar{X}}_{98.ag})^2}{S_{ag} - 1}\right)^{1/2}. \tag{2.3}$$

Second, an adjusted 1998 z score for every school/grade by subject was computed by multiplying the respective z_{98sag} by the correlation between scaled scores of adjoining grades for all schools. The correlation ($\hat{r}_{98,99\,ag}$) was computed at the student level. The equation for the adjustment is

$$\tilde{z}_{98sag} = z_{98sag} * \hat{r}_{98,99\,ag}. \tag{2.4}$$

Third, an adjusted 1998 scaled score was then computed by transforming the adjusted z score back into the metric of the scaled score. The equation for this transformation is

$$\tilde{\bar{x}}_{98sag} = (\tilde{z}_{98sag} * \hat{\sigma}_{98.ag}) + \bar{\bar{x}}_{98.ag}. \tag{2.5}$$

Fourth, the adjusted growth was then computed by subtracting the adjusted 1998 scaled score from the 1999 scaled score. The equation for adjusted growth is

$$AG_{98sag} = \bar{x}_{99sag} - \tilde{\bar{x}}_{98sag}. \tag{2.6}$$

Unit of analysis. The opening sentence of Arizona's first MAP report states, "The Arizona Measure of Academic Progress (MAP) has been developed to examine the academic growth of individual students over time." The method for reaching this goal started with achievement scores matched by students across two consecutive years. Instead of computing growth at the student level, the means for each school/grade unit were computed. Growth was then computed as a function of the means of each school/grade unit. Individual student growth was never computed nor does the report examine individual growth.

The level at which student learning is computed is important because additional information can be gained by computing growth at the individual student level. In practice, the results from individual students could be combined for analyses of unit levels other than school/grade. For example, students from a single class could be used as part of a teacher-level assessment. Arizona's method could have started by computing growth for each student and then averaging the growth within school/grade units.

Scaled scores. There are two issues related to the scaled scores that make their use suspect. Before these issues are explored, a general definition of scaled scores is provided. Scaled scores are the values assigned to students on the basis of test performance that are intended to reflect levels of achievement (Petersen, Kolen, & Hoover, 1993).

Achievement tests such as the Stanford 9 are administered as separate forms based on the grade level of the student. When test builders want a common scale across all grade levels, they vertically equate the scores from all the forms of the test to

construct a scale continuum across all the grade levels. For example the Stanford 9 has scaled scores that range from approximately 400 for low achieving second and third graders to 800 for high achieving eighth and ninth graders (the actual limits of the Stanford 9 scaled scores from grades K though 12 are 100 to 999).

The specific process of how scaled scores were produced for the Stanford 9 includes three main components. The first is the estimation of item difficulties across the different grade-level forms. The second component is to produce a common scale across the multiple forms in a process called vertical equating. The last component is to re-scale the common scores to produce a more convenient scale for teachers and parents to interpret.

Item difficulties for questions in the Stanford 9 were estimated using an Item Response Theory (IRT) model, specifically the one-parameter Rasch model (Lord, 1980). The Rasch model was chosen mainly because it facilitates the construction of a common scale across different ability levels and also because it assumes a one-to-one relationship between raw scores and ability estimates (i.e., the ability estimate is the same for all individuals with the same raw score on a test regardless of the particular items answered correctly). This latter feature makes it possible for scaled scores to be assigned by test users who must hand score the exam.

To place scores for all grade levels on the same scale, vertical equating was carried out by administering two adjacent test levels to an equating sample. The item difficulties were estimated using the Rasch model for both levels treated as one long test. The average item difficulty was then computed for each level. The difference between the

average item difficulty of a common subtest between two levels was determined and used as an equating constant to convert the item difficulties of one test level to the scale of the next test level. This process was conducted for all adjacent levels of the Stanford 9 and for each content domain.

The last step in constructing the scaled scores was to produce a convenient scale. A linear transformation was performed resulting in a three-digit integer scale. The main benefit of this transformation was to produce a scale represented by positive integers.

Problems with scaled scores. The Rasch model is a special case of a more general three-parameter logistic model. In the univariate model, it is assumed that there is a single ability underlying all the items in the test (Lord, 1980). In addition, local independence is assumed, that is, for a subject of a given ability, the answers to different items are independent. The probability that a student with ability θ will answer an item correctly is

$$P_i(\theta_j) = c_i + (1 - c_i) \frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \qquad (2.7)$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the guessing parameter for item i, and D is a scaling factor of 1.7. The Rasch model assumes that the guessing parameters are zero and that all discrimination parameters are equal. Of the item parameters, only the difficulties vary in the Rasch model. A feature of the model is that the raw number-correct score on a test is sufficient for estimating ability.

The construction of scaled scores that incorporate multiple forms (for different achievement levels) is typically referred to as vertical equating (Petersen, Kolen, & Hoover, 1993). There are problems that surface in the vertical equating process when the

Rasch model is used with multiple-choice items. Research by Divgi (1986) has demonstrated that the Rasch model lacks sufficient information needed to model adequately the responses from multiple-choice tests. Essentially, this research showed that when the Rasch model is used with multiple-choice items, a large proportion of items do not fit the model. When the model fails to fit a large proportion of items, the properties of the test are different from those of the model. In particular, ability estimates based on the model will not be "item free." If the Rasch model does not fit the data, then the resulting scale (and subsequent vertical scale) will not adequately reflect the true developmental continuum.

Slinde and Linn (1979) found problems when the Rasch model was used to equate different difficulty levels of tests. When the average difficulty of the items between the forms was similar (as in horizontal equating), the Rasch-based equating appeared to function properly. When the difference between item difficulties grew between the two forms (as in vertical equating), the model did not fit. Slinde and Linn believed that guessing was a major contributor to the fit problems.

The problems with the use of scaled scores for the analysis of growth are important to consider. Given that Stanford 9 scaled scores are based on the Rasch model, it is possible they do not adequately reflect the content knowledge they are designed to measure. An additional issue with the scaled scores and the OYG indicator in the Arizona method relates to the comparison of developmental scale scores to norm-referenced indicators. This issue are described in greater detail in a discussion of OYG.

Regression to the mean. It is argued that the correction for the effects of RTM is inappropriate for this particular analysis. Campbell and Kenny (1999) note that the phenomenon of RTM can be subtle and easy to miss and that the concept is widely misunderstood. For this reason, this section starts with two definitions of RTM.

Cook and Campbell (1979) provide a conceptual description of RTM in their discussions of threats to internal validity when they are talking about "statistical regression." When participants are classified into an experimental group on the basis of their pretest scores or some correlate of the scores and the measures are not perfectly reliable, high pretest scorers will score relatively lower at the post-test and low pretest scorers will score higher at the post-test. This tendency occurs even if no treatment is administered. Thus, the change in group means in an experiment from pre to post may be due in part, or whole, to a statistical artifact. It is for this reason that past researchers have suggested corrections for RTM. By applying a correction for RTM, the experimental effects can be better estimated.

The part of Cook and Campbell's definition that is relevant to Arizona's method is the manner of classifying participants into experimental groups. If participants are not classified into experimental groups based on pretest scores, or some correlate of pretest scores, RTM does not occur. Other definitions of RTM are required to further explain why the selection criteria is important.

A more technical definition of RTM is provided from Lund (1989). For a given population, statistical regression is defined as the difference between the predicted standard score in one variable and the selected standard score in the other variable.

Suppose that X and Y are two random variables representing a pre and post measure. If Y is predicted from X, the regression effect will be

$$Z'_Y - Z_X \tag{2.8}$$

where Z'_Y is the predicted Y score expressed in standard score units and Z_X is the selected X score also expressed in standard score units. The standard scores are given by the regression equation for standard scores

$$Z'_Y = r_{XY}(Z_X) \tag{2.9}$$

where r_{XY} is the correlation between X and Y. For example if $Z_X = 2.0$ and $r_{XY} = .80$ then Z'_Y would equal 1.6. If $Z_X = -.5$ and $r_{XY} = .80$ then Z'_Y would equal -.4. The closer the absolute value of the X-selected score is to the mean of X the smaller the effect of RTM will be.

Lund's definition for RTM involves two assumptions. The first is that the population of subjects is completely specified, meaning that the sample adequately describes the population. If the population is incompletely specified, the regression effect is not accurately estimated. The second assumption, and more critical to Arizona's method, is that of direct selection of students on one of the variables.

Choosing the highest five scores from the X variable would be an example of direct selection. Indirect selection occurs when some variable other than X or Y is used to select cases from X. (There are trivial cases where the external variable is perfectly correlated with X or zero-correlated with X.) The relationship between the external variable used to select from X and the X variable is usually unknown. When external selection is used, there is not a systematic way to predict what the means of Y will be by

using the standard RTM equations. When an external variable is used to select from *X*, the means for those groups on the *Y* variable could go up or down or stay the same. The change in the Y variable is no longer a simple function of the correlation between X and Y. Therefore it is not an effect of RTM as given by the definition.

In the Arizona method, the students are being indirectly selected using the external variable "school membership." Thus, the prediction of the post-mean for a particular school from its pre-mean is no longer a simple linear function. Therefore, using a standard correction for RTM when an external variable is used for selection is not appropriate. The Dallas Value-Added Accountability System corrects for what Webster (1996) calls "slippage". Rather than applying a standard correction for RTM, the HLM model used in the Dallas system includes more variables than just the correlation between the pre and post measures. The additional variables are likely to produce a more accurate correction for extraneous uncontrolled variables over an RTM correction based on Lund's definition.

Inappropriately correcting for RTM can lead to dismissing real educational effects as statistical artifacts. The correction systematically leads to reducing the amount of growth needed to maintain high scores while increasing the amount of growth needed for low scorers. If the amount of growth required is unwarranted, then the correction is certainly unfair to lower-scoring students and overly generous to higher-scoring students.

One-Year's Growth in the Arizona Method

The third step in the Arizona method was the computation of the OYG indicator for each school/grade unit by subject area. The decision of whether a unit had achieved

OYG was based on a comparison of the observed amount of growth (the adjusted growth) and an expected amount. The expected amount of growth came from the Stanford 9 Technical Data Report (1997). It was the scaled score change from one year to the next, based on the national norm sample. Specifically, the amount of growth (in scaled score points) made by a student performing at the median was used. The expected amount of growth at the median was chosen because it corresponds to the part of the test with the least amount of measurement error under the Rasch model.

If the adjusted growth was equal to or larger than the expected amount, then the school/grade unit achieved OYG in that subject. If the adjusted growth was less than expected, the school/grade unit did not achieve OYG. The dichotomous OYG decision was reported for each school/grade by subject area in the state. In the case of Reading from grades 4 to 5 for School A,

$$OYG_{ARead4,5} = \text{"Yes"} \text{ if } AdjustedGrowth_{ARead4,5} >= EG_{Read4,5}$$

$$OYG_{ARead4,5} = \text{"No"} \text{ if } AdjustedGrowth_{ARead4,5} < EG_{Read4,5}$$

The OYG indicator. OYG is a norm-referenced indicator. The concept of OYG is asking the question: "Did the students learn as much as some other group with similar characteristics learned over the course of a year?" If one assumes the average national growth to be "normal growth," then the concept of OYG can be used to determine if normal academic growth is occurring. In the case of the Arizona method, researchers wanted to know if the average amount of student growth that took place in a school was equivalent to the average amount of growth that took place in the nation (the norm

sample of the Stanford 9). OYG is not an absolute measure of academic growth. Therefore the amount of learning that took place from one year to the next is not known.

In the Arizona method, OYG is computed at the school/grade level. However, OYG could be computed for each individual student. In either case, there are different ways to approach the question. One approach could use percentile ranking. Suppose a student has percentile ranking of 60 in Mathematics at the end of grade 4. If that student grew as much, during the 5th grade, as others in the nation, she would have a percentile ranking of 60 the next year. There is another approach that could be used if a score type were chosen which is not in a norm-referenced metric (such as scaled scores). The amount of growth needed to maintain the same relative ranking for that particular type of score could be determined. Some technical issues are related to the latter approach and are detailed using the Arizona method as an example.

Scaled scores require a norm-referenced expected growth value to assess whether students maintain the same national percentile ranking from one year to the next. The expected growth values are different for students performing at different ability levels. In the Arizona method, the expected growth for only the 50th percentile is used to make judgments concerning academic growth for students at every percentile rank.

Tables 2.2 and 2.3 show the expected growth, in scaled score points, needed for students at key percentile ranks to maintain those ranks from one grade level to the next.

Table 2.2

Expected Scaled Score Growth at Key Percentile Rankings (PR), Mathematics

	Expected Growth Between Grades				
PR	3-4	4-5	5-6	6-7	7-8
P90	23	15	20	10	10
P75	25	21	11	14	11
P50	26	21	10	14	10
P25	27	25	7	17	9
P10	27	25	8	18	7

Table 2.3

Expected Scaled Score Growth at Key Percentile Rankings (PR), Reading

	Expected Growth Between Grades				
PR	3-4	4-5	5-6	6-7	7-8
P90	15	14	7	15	17
P75	21	12	8	18	11
P50	21	18	8	18	10
P25	23	19	9	16	12
P10	24	19	12	13	12

It is evident from Tables 2.2 and 2.3 that within a grade level change, the number of scaled score points needed to make OYG differ over percentile ranks. In effect, scaled scores are not equal interval with respect to one year's worth of academic growth. Because of this problem, it is inappropriate to compare the growth values of students at a particular percentile rank with those at any other percentile rank, even if they are in the

same grades. Computing the scaled score growth mean for a school/grade unit is misleading because the average is computed over intervals which are not equivalent in terms of what they represent.

The Star Rating in the Arizona Method

The last step of the Arizona method was the computation of the Star Rating for each school/grade unit. The Star Rating is a ranking indicator that allows schools to compare themselves with other schools in Arizona. To compute the Star Rating, the adjusted growth values for all schools at the same grade levels were ranked and quintiles were produced. A number of "Stars" were assigned to each school/grade unit based on their quintile ranging from 1 (Low) to 5 (Excellent). This indicator was called the Star Rating. In Arizona's case researchers chose a five-point scale because they felt the metric would be publicly consumable. The number of points on the scale is of course subjective.

Working only with a norm-referenced indicator like OYG and a within-state rank indicator like the Star Rating may seem limiting, but a good deal of useful information can be gleaned. The Star Rating and the OYG indicators are independent of each other. For example, a third grade in School A can have a Star Rating of 2 (below average) and still achieve OYG. The indicators inform the principal of the school that although her third grade grew at the same or higher rate of the nation, it grew at a slower rate than most other third grades in the state.

Arizona's initial method for estimating school effectiveness from student achievement scores has problems. The scaled scores contribute to a large number of the issues. Even with the problems, there are many positive features of the method. For

example, the matched student data and the selection criteria for students and school/grade units add to the internal validity of the method by ensuring equivalent pre-to-post samples. Having a norm-referenced indicator (OYG) in addition to a within-state rank-ordered indicator (Star Rating) provides useful information to the schools.

Part II: Alternative Methods

The approach Arizona chose for the first MAP analysis was just one of several routes by which academic growth could have been estimated. Using some of the better features of the Arizona method, as well as other metrics and procedures, three alternative approaches are suggested.

Simple Growth

All of the alternative methods make use of simple growth. To begin, the rationale for the use of simple growth and some related issues are presented. The simple difference between two observed scores for a student can be used to estimate the amount of growth that the student has made over the course of an academic year in the measure under consideration. For educational researchers, the computation is simple and the change scores from multiple students can be used as a dependent variable. The advantage of change scores is that it is possible to compare the changes of individuals or groups that started with unequal pretest scores. There has been considerable debate over the appropriateness of using simple growth for two primary reasons. First, the reliabilities of change scores are typically low. Second, there is a systematic relationship between measures of change and initial status (Feldt & Brennan, 1993). Linn (1981) explains that

the low reliability of difference scores is a more serious problem when used to make decisions about individuals. For groups of students, the reliability is of less concern and is not a fatal flaw in educational research.

The systematic relationship between measures of change and initial status can have a greater effect in educational research. When pretest and posttest variances are similar, the change scores will tend to correlate negatively with initial status. Students with low pretest scores will tend to have larger growth. If the posttest variance is greater than the pretest variance, the change scores will tend to correlate positively with initial status. Students with high pretest scores will tend to have larger growth (Linn 1981).

Normal Curve Equivalent (NCE) Scores

The NCE score was introduced for use in the Title I Evaluation and Reporting System. NCEs are normalized (scores expressed in terms of a normal distribution) standard scores with a mean of 50 and a standard deviation of 21.06, ranging from 1 to 99 with the same numerical value as a percentile rank at scores of 1, 50, and 99. At other scale points the two types of scores do not correspond. Under certain assumptions, NCEs are considered to have an equal-interval scale, allowing the computation of aggregated or averaged NCE scores. The NCE scale is essentially the same as a normalized z score with a different mean and standard deviation.

Alternative Method 1

The first alternative method takes advantage of the equal-interval nature of NCE scores to compute the amount of academic growth. Starting with the basis of the Arizona method, the student matching criteria as well as the inclusion requirements for

school/grade units are employed. The student matching ensures that the variations in the types of students from one year to the next are held constant within each school/grade unit. The school/grade inclusion requirements exclude units that are too small to allow valid conclusions.

Rather than aggregating scores within a school/grade unit, academic growth is assessed by computing the difference between the NCE scores for each student in the matched sample. The NCE difference gives each student a measure of growth that is relative to the normative population. The relative nature of the difference is based on the assumption that, under normal conditions, a student should be expected to maintain the same relative standing from one testing to the next. A student who receives the same or higher NCE score for two consecutive years would have made at least one year's worth of academic growth. Computing growth at the student level first creates more information that can be used in later analyses.

The growth mean for a school/grade unit is computed by finding the mean of the differences of the individual students' NCE scores across the two years in the matched data. For the reasons previously mentioned, the growth for each school/grade unit does not need to be corrected for RTM. The simple growth mean can be used to compute the academic growth indicators reported to each school.

To compute an OYG decision for each school/grade unit, the simple growth mean can be dichotomized. Because NCE scores are used, school/grade units would achieve OYG if the growth mean is zero or higher. In the case of Reading from grades 3 to 4 for School A,

$$OYG_{ARead3,4} = \text{"Yes"} \text{ if } \bar{x}_{GrowthARead3,4} >= 0$$

$$OYG_{ARead3,4} = \text{"No"} \text{ if } \bar{x}_{GrowthARead3,4} < 0$$

A within-groups rank-ordered comparison similar to the "Star Rating" of the Arizona method can be computed from the simple growth mean as well. To accomplish rankings, the simple growth mean for a particular grade level can be rank ordered across all the schools within the state. From the within-state rankings, quintiles can be assigned to each school/grade unit.

Dichotomized Growth

In the Arizona method and in the first alternative method growth is represented as a function of "amount of change." This amount is based on the difference of two scores. Growth can also be expressed in a simpler mode of "yes" (it occurred), or "no" (it did not). For example, one can determine whether a student had achieved one year of growth. While such an approach has a reduced amount of information, there are advantages as well. To compute "amount of growth," an equal-interval scale is required. To compute a dichotomized or directional growth, the scales only need to meet an ordinal requirement. In addition, scales with higher orders of measurement can be used to compute dichotomized growth when the assumptions for the scale may be in question. For example, it may be the case that the first alternative method cannot be used because assumptions have been violated which are necessary for the equal-interval scale of the NCE scores. Based on possibilities like assumption violations, a second alternative method is proposed.

Alternative Method 2 with NCE Scores

The second alternative method would also start with the same basis of the Arizona method, that of the matched student criteria as well as the minimum size requirements for a school/grade unit. Again, rather than aggregating scores within a school/grade unit, academic growth can be assessed at the student level. Making use of the NCE score and the notion of dichotomized growth, a growth decision can be computed for each student by comparing the NCE scores of successive years. If a student receives the same or higher NCE the second year, then that student has made $OYG^{Student}$. Like the first alternative, this method would not need a correction for RTM.

To compute an OYG decision for a particular school/grade unit, the proportion of students who achieved $OYG^{Student}$ can be computed within that unit. A criterion can then be used to make the decision. For example, if at least half the students achieve $OYG^{Student}$, then the school/grade unit also achieves OYG. In the case of Reading from grades 3 to 4 for School A,

$$OYG_{ARead3,4} = \text{"Yes"} \text{ if proportion of } OYG^{Student}_{ARead3,4} >= .50$$

$$OYG_{ARead3,4} = \text{"No"} \text{ if proportion of } OYG^{Student}_{ARead3,4} < .50$$

Again, a within-groups rank-ordered comparison similar to the "Star Rating" of the Arizona method can be computed. To accomplish rankings, the proportion of $OYG^{Student}$ for a particular grade level can be rank-ordered across all the schools within the state. From the within-state rankings, quintiles can be assigned to each school/grade unit.

Percentile Rank Scores

A percentile rank (PR) score provides information about a student's relative standing in comparison to a norm group. PR scores range from a low of 1 to a high of 99, with 50 indicating average performance. PR scores can be computed from lower-, mid-, or upper-interval cumulative percentage distributions. On the Stanford 9, mid-interval PR scores are computed. The percentile rank score corresponds to the percentage of the norm group obtaining scores equal to or less than that score. For example, a student with a PR score of 55 performed as well or better than 55% of the other students in the norm group. While PR scores are ordinal, they do not have equal intervals, so they cannot be aggregated or averaged.

Alternative Method 2 with PR Scores

As mentioned before, one of the advantages of computing a dichotomized growth indicator is that the scales only need to meet an ordinal requirement. Thus PR scores can be used for this purpose. A variation of the second alternative method could use PR scores rather than NCE scores. PR scores may be desirable if it were the case that NCE scores were not available. In the variation of the second alternative, all of the procedures are exactly the same, with PR scores substituted for NCE scores.

The dichotomized growth decision can be computed for each student by comparing the PR scores of successive years. If a student receives the same or higher PR the second year, then that student has made $OYG^{Student}$.

To compute an OYG decision for a particular school/grade unit, the proportion of students who achieved $OYG^{Student}$ can be computed within that unit. Again, a criterion of

50% can then be used to make the decision. Finally, the "Star Rating" can be computed from the proportion of students achieving one year's growth within a school/grade unit.

The results of the second alternative method would not be affected by the use of NCE or PR scores. While NCE and PR scores correspond only at the values of 1, 50, and 99, either approach would lead to the same OYG decision for each student. The remaining calculations and results at the school/grade level of analysis would be identical regardless of the score used.

Accounting for Error

While the reliabilities of standardized achievement tests are typically high, they do not measure students without error. According to classical test theory, a student's observed score is a function of her true score and error. When an imperfect measuring device is used, the result is an imperfect estimate of true ability.

The amount of error associated with a point estimate can be incorporated by using a confidence interval. Confidence intervals are bands surrounding an estimate that represent a specified probability that the true value is located somewhere within the bands. A confidence interval for an observed value x_t may be defined as

$$CI_{OBS} = x_t \pm z_{\alpha/2}(\sigma_e) \tag{2.10}$$

where $z_{\alpha/2}$ is the standard unit associated with a desired probability, and σ_e is the standard error of the observed score. When considering a single score, the width of the confidence interval is a function of the desired confidence level and the standard error of measurement (σ_e). The classic equation for σ_e is based on standard deviation and the reliability of the test.

$$\sigma_e = \sigma_{x_t}\sqrt{1-\rho_{tt}} \tag{2.11}$$

where σ_{x_t} is the standard deviation of the observed scores and ρ_{tt} is the reliability of test *t*.

Arizona's method did not take the error associated with students' scores into account. Critics from schools in Arizona pointed out that it may not be fair to say a student did not achieve OYG simply because she grew one point short of the expected value. Error can be accounted for in the method, but the approaches are not straightforward.

The equation for the SEM above can be viewed as the expected value of the person-specific standard errors of measurement (PSEMs) that could be estimated were it possible to conduct a repeated-measure experiment on every examinee selected (Traub, 1994). The standard errors of scores are not equal across different ability levels. Standardized achievement tests represent a competitive environment where cutoff scores have serious consequences. Building confidence intervals around scores using the same SEM at different levels of abilities would yield inaccurate results.

There are procedures for estimating the standard errors at the individual score level; however, they are impractical to pursue at the state-level. For example, the Stanford 9 Technical Data Report provides only reliability estimates for a particular form of a test. SEM values are not given for specific scores. It would be difficult at the state-level to compute accurate confidence intervals around the scores of students. To make matters even more complicated, the measurement of growth is concerned with the

difference between scores, a value that has a different standard error than for a single score.

There are simpler ways to account for error in the measurement of academic growth. Rather than working a confidence interval into the score differences at the student level, standard errors can be incorporated at the school/grade level.

In the Arizona method, as well as the two alternative methods already proposed, the OYG indicator is computed for each school/grade unit. Passing or failure is a critical decision to be making for a school principal. If the average growth of a school/grade unit falls just short of what was expected, then that principal will need to be convinced that the "No" decision for OYG is justified.

In the first alternative method proposed, the growth for a school/grade unit is the mean of the differences of the students' NCE scores across the two years of matched data. Instead of relying on measurement error at the student level, the variation of student growth can be used at the school/grade level. The standard error of the students' score differences within a school/grade unit would be computed as

$$\hat{\sigma}_{\bar{x}sag} = \frac{\hat{\sigma}_{xsag}}{\sqrt{n_{sag}}} \tag{2.12}$$

where $\hat{\sigma}_{xsag}$ is the standard deviation of the growth mean and n_{sag} is the number of students for school s, in area a, and grade g. The standard error would be added to the observed growth mean for each school/grade unit to produce a growth mean adjusted to account for error. This adjusted value would then be compared to the OYG criteria.

Using the approach above would result in different standard errors for school/grade units. Unfortunately, holding schools to different standards would be difficult to defend. Two schools with the same amount of variability, but different numbers of students, would have different adjustments to account for error. Rather than computing a standard error for each school/grade unit, a single standard error for each grade level can be computed based on the total variability of student growth within schools and the average unit size in each grade level. The equation would be

$$\hat{\sigma}_{\bar{x}ag} = \sqrt{\frac{\sum_{s=1}^{Sag}\left(\sum_{i=1}^{Nsag}(x_{isag} - \bar{x}_{.sag})^2\right)}{(N_{ag} - S_{ag})(\bar{N}_{.ag})}} \qquad (2.13)$$

where x_{isag} is the NCE growth made by student i in school s in area a in grade g, $\bar{x}_{.sag}$ is the NCE growth mean for school s in area a in grade g, N_{sag} is the number of students in area a in grade g, S_{ag} is the number of schools in area a in grade g, and $\bar{N}_{.ag}$ is the average number of students in schools in area a in grade g.

This grade level standard error would be subtracted from the OYG criterion to produce an adjusted criterion to account for error. The observed growth mean would then be compared to the adjusted criterion to make an OYG decision for each school/grade unit.

In the second alternative method mentioned, the OYG decision is based on the proportion of students achieving OYG for each school/grade unit. Accounting for error in the decision can be accomplished by estimating the error associated with the proportion of students achieving OYG in a school/grade unit. This approach would also result with a

different standard error amount for each school in a grade. For the reasons mentioned above, a single standard error can be constructed for each grade level. To result in a common standard error of proportion for each grade level, the standard error would be based on the proportion criterion, rather the proportion in each school/grade unit. In the second alternative method, the criterion for a school/grade unit to achieve OYG is that 50% of students need to achieve OYG. The following is an equation for a standard error of proportion based on the criteria of 50% and the average unit size for each grade level:

$$\hat{\sigma}_{Pag} = \sqrt{\frac{.50(1-.50)}{\overline{N}_{.ag}}} \qquad (2.14)$$

where $\overline{N}_{.ag}$ is the average number of students within a school in area a in grade g.

To use this standard error, the criterion of .50 in the second alternative method would be reduced for each grade level by $\hat{\sigma}_{Pag}$. The OYG decision for school/grade units would then be based on exceeding the adjusted criterion.

Stanine Scores

Stanines are scores expressed in terms of a distribution that has been transformed to fit a normal curve. Stanines, developed by the United States Air Force during World War II, provide a single-digit system of scores with a mean of 5 and standard deviation of approximately 2, and range from a low of 1 to a high of 9. Like PR scores, stanines are ordinal but not equal-interval (Anastasi, 1988).

A disadvantage with stanines is that there are only nine scores possible on the entire scale, yielding only rough estimates of student ability. This disadvantage can be turned into a benefit if the stanines are used to account for error when making the OYG

decisions. For a majority of students, correctly answering one or two additional items on a test will not affect their stanine scores. This would not be the case for NCE or PR scores, as one additional correct question can change the score. Stanines are certainly a function of the number correct, but they are less sensitive to minor score variations than other score types. The stability of a student's stanine score works in favor of those wishing to account for measurement error because the stanine will generally not fluctuate as much as other scores types, given the same amount of error.

Alternative Method 3

The third alternative method also starts with the same basis for the Arizona method (using the matched student criteria as well as the inclusion criteria for school/grade units). Once again, rather than aggregating scores within a school/grade unit, academic growth is assessed at the student level. Making use of the stanine score and the notion of dichotomized growth, a growth decision can be computed for each student by comparing the stanine scores of successive years. If a student receives the same or higher stanine the second year, then that student has made $OYG^{Student}$. To compute an OYG decision for a particular school/grade unit, the proportion of $OYG^{Student}$ can be computed. A criterion of 50% can be used to make the decision.

Finally, a within-groups rank-ordered comparison similar to the "Star Rating" of the Arizona method can be computed from the proportion of $OYG^{Student}$ within a school/grade unit. To compute the Star Rating the proportion of $OYG^{Student}$ for a particular grade level can be rank-ordered across all the schools within the state. From the within-state rankings, quintiles can be assigned to each school/grade unit.

Part III: The Current Research Project

It is important that the Arizona Department of Education chooses a method that is based on sound principles, useful to those it is meant to serve, and represents a valid measure of student growth. The particular method chosen is important because the results of the analyses may vary from one method to another.

The initial method used by Arizona has some aspects which are sound, but others which are not. Three alternative methods to assess the performance of school/grade units using student academic growth have been presented. In order to demonstrate the effect of method choice, this research compares the Arizona method to the three alternative methods. To see how the methods would actually perform in practice, the scores from students in Arizona were used in re-computing measures of growth as specified in Part II. Comparisons are made from several perspectives because there are several key features that define the differences among the methods.

Category I:

The first set of research questions examines the impact of the method choices in Arizona's approach. First, the OYG decision in Arizona's method may be dependent on grade level because expected growth differs from one grade level to the next. Second, due to the differences between the methods, the same school/grade unit may or may not achieve OYG. Third, the methods define growth differently and may differ in sensitivity to extreme values. Thus, the impact of outliers will be greater when they occur in school/grade units with a small number of students, thereby differentially changing the

results. Finally, due to the differences among the methods, the same school/grade unit may or may not receive the same Star Rating.

To draw better generalizations from the results, only the methods that do not account for measurement error were used. In particular, the Arizona method and the first and second alternative methods were compared in terms of the OYG decision and Star Rating for each school/grade unit. The questions that needed to be addressed in this category were:

I-1: What are the differences among Methods AZ, A1, and A2 in the assignment of OYG to schools across grade levels for Reading and Mathematics?

I-2: Do the OYG decisions of Arizona's method and alternative methods 1 and 2 agree for each unit across grade levels for Reading and Mathematics?

I-3: Does the size of the school/grade unit have an effect on the agreement of OYG between Arizona's method and alternative methods 1 and 2 for Reading and Mathematics?

I-4: Do the results differ between Arizona's method and alternative methods 1 and 2 for computing Growth Rankings (The Star Rating) for Reading and Mathematics?

Category II:

The second set of research questions examined the impact of Arizona's correction for RTM on the OYG decision and Star Rating for each school/grade unit. First, because of the correction, the same school/grade unit may or may not achieve OYG. Second, in addition to the OYG indicator, the Star Rating may be influenced by the correction for

RTM. Third, the impact of correcting for RTM will be greater on units whose initial means are far above or below the overall mean for all units within a grade level.

To isolate the effects of correction for RTM, the Arizona method was compared with and without the correction. The questions for this category were:

II-1: Do the OYG decisions of Arizona's methods (with and without a correction for RTM) agree for each school/grade unit for Reading and Mathematics?

II-2: Do Arizona's methods (with and without a correction for RTM) give different Growth Rankings (Star Rating) for Reading and Mathematics?

II-3: Does the initial scaled score mean for the school/grade units affect the agreement for the OYG decision between Arizona's methods (with and without a correction for RTM) for Reading and Mathematics?

Category III:

The third set of research questions examined the independence of the change indicators with initial status of the school/grade unit. Ideally, a growth indicator should not be dependent on the initial status (the pre-measure mean for the unit). The greater the independence, the more specific the information will be. This is an important question for Arizona's method as well as all three of the alternative methods proposed. If the *amount of growth* (the simple or adjusted growth) for a school/grade unit is uncorrelated with initial status, the indictor is providing unique information. If the OYG decision for a school/grade unit is uncorrelated with initial status, that indictor is also providing unique information. The questions in this category that needed to be answered were:

III-1: What are the correlations between the amount of growth and initial status of the school/grade unit for Arizona's method and alternative methods 1, 2, and 3?

III-2: What are the correlations between the OYG decision and initial status of the school/grade unit for Arizona's method and alternative methods 1, 2, and 3?

Category IV:

The fourth set of research questions examined the effect of accounting for error in the three alternative methods on the OYG decision for each school/grade unit. First, making the adjustments to account for error in the first and second alternative methods will increase the number of school/grade units achieving OYG. The degree to which it will do so is unknown. Second, the adjustments to account for error in the first two alternative methods are conceptually different from the manner in which the third alternative method accounts for error. The similarities and differences between these methods are unknown.

To make certain that the adjustments are the only influence on alternative methods 1 and 2, they were compared with and without adjustments. Finally, the differences between alternative methods 1 and 2 with the adjustments and alternative method 3 are examined. The research questions for this category were:

IV-1: Do the OYG decisions of the first alternative method, with and without adjustments to account for error, agree for each school/grade unit for Reading and Mathematics?

IV-2: Do the OYG decisions of the second alternative method, with and without adjustments to account for error, agree for each school/grade unit for Reading and Mathematics?

IV-3: Do the OYG decisions of the alternative methods 1 and 2, with adjustments to account for error, and alternative method 3 agree for each school/grade unit for Reading and Mathematics?

Category V:

The fifth and final category has only one research question. It was used to determine the minimum number of students in a school/grade unit necessary to provide acceptable accuracy in estimating academic growth of each unit. This question was explored in the context of minimizing the standard error of the estimates and the proportion of school/grade units that would be omitted from analyses.

This research question estimates the number of school/grade units in Arizona that would be omitted from future analyses given varying minimum size criteria. The research question was:

V-1: What proportion of school/grade units are excluded for different minimum size criteria?

CHAPTER 3

PROCEDURES

To answer the research questions discussed in the previous chapter, the results of calculations made using Arizona's method and the three alternative methods were compared in a variety of conditions. Arizona's method was replicated exactly as it was done for the first MAP report. In this research, Arizona's method is referred to as Method AZ. The three alternative methods suggested in the previous chapter are referred to as Methods A1, A2, and A3. The criteria used to include students and school/grade units are detailed in Appendix A.

The Student Achievement Data

The initial database from the Arizona Department of Education contained the scores for every third through seventh grade student in Arizona who took the Stanford 9 Achievement tests in 1998 and every fourth though eighth grade student in 1999. In 1998 there were approximately 410,000 students represented. In 1999 there were approximately 450,000 students. The variables included in the data were: the grade levels of each student in 1999 and 1998, school identification, student identification, scaled scores, percentile rank scores, and normal curve equivalent scores for Reading, Mathematics, and Language for 1999 and 1998. The Language scores were not analyzed because the Department of Education in the State of Arizona summarizes only Reading and Mathematics. Student-record matching variables were also included to reproduce, as

closely as possible, the matched data set Arizona used for its report. To maintain confidentiality, school names and students identification were re-coded.

The selection criteria for students and school/grade units were kept the same as the State's method (see Appendix A for a complete listing of the criteria). Table 3.1 contains summary information on the sample used for the analyses.

Table 3.1

Number of School/Grade Units and Students

	Reading		Math	
Grade Level	Number of Units	Number of Students	Number of Units	Number of Students
3-4	643	35,109	648	36,448
4-5	659	37,572	661	38,863
5-6	445	26,187	449	26,794
6-7	256	20,611	256	20,853
7-8	332	40,227	332	40,106
Total	2,335	159,706	2,346	163,064

Analysis of Category I:

The Differences among Methods AZ, A1, and A2

The first set of research questions examined the differences among the methods. The four research questions in this category focused on the differences and similarities among methods in the assignment of OYG to school/grade units as well as the growth rankings (the Star Ratings). The analysis procedures for each research question are listed separately.

The matched-student Stanford 9 achievement data were used to address all four of the questions in this category. The analyses for questions I - 1, I - 2, and I - 3 began by computing the OYG decision separately for each school/grade unit using Methods AZ, A1, and A2. The three analyses resulted in a dataset with three dichotomous indicators of OYG for every school/grade unit. For question I - 4, Star Ratings were computed for each school/grade unit using Methods AZ, A1, and A2. Six variables (the three Star Ratings and three OYG decisions) for each school/grade unit were used for comparing the methods. All analyses were completed separately for Reading and Mathematics.

Analysis of Research Question I - 1

Research question I - 1: What are the differences among Methods AZ, A1, and A2 in the assignment of OYG to schools across grade levels for Reading and Mathematics? To answer this question, log-linear analyses were used to determine whether the proportion of schools achieving OYG depended on method choice (AZ, A1, or A2) and grade level.

Log-linear analysis is a method for analyzing the frequency of observations in each cell of a multi-way contingency table. It has features similar to an ANOVA effects model in which the number of observations in each cell is a function of main and interaction effects. In a log-linear analysis, each cell (or each cross-classification of the independent variables) is expressed in terms of observed probability of occurrence by dividing the number of observations in the cell by the total number of observations in the analysis. Multiplicative procedures can be used to model the probabilities of cell occurrence; however, by taking the natural logs of the cell probabilities, the models can be expressed as a linear function. These "log-linear" models result in effect parameters that are interpreted in a similar manner as the effects from an ANOVA model (Kennedy, 1983). The log probabilities for each cell of a contingency table are estimated using a log-linear model containing the effects thought to contribute to the observed data in the cells. If a log-linear model contains parameters for the proper effects, the model will "adequately" (in a statistical sense) reproduce the observed data.

Haberman (1979) describes different model classifications based on the number and types of effects included in the model. A model that contains the main effects and all higher-order interaction effects of the independent variables is a "saturated model". A saturated log-linear model contains as many effects as cells. For example, in a 2 x 2 contingency table the saturated log-linear model is given as

$$\upsilon_{ij} = \overline{\upsilon}.. + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \qquad (3.1)$$

where υ_{ij} is the log-probability of an observation in cell (i, j). $\overline{\upsilon}..$ is the mean log-probability of all the cells; it is analogous to the grand mean of an ANOVA effects

model. λ_i^A and λ_j^B are the model parameters for the main effects of the i^{th} level of variable A and the j^{th} level of variable B in log-probability units. Because of the constraints on the model, namely that the effects sum to zero, a – 1 independent parameters can be estimated for effect A, and b – 1 independent parameters can be estimated for effect B. λ_{ij}^{AB} is the model parameter for the interaction effect of the i^{th} level of A and the j^{th} level of B in log-probability units. There are (a - 1)(b - 1) independent parameters estimated for an interaction effect in a log-linear model because of linear constraints. If $\lambda_i^A = 0$ for all values of i, then there are no main effects of variable A, meaning that each level of A is equally likely. If $\lambda_i^A > 0$, then the first category of A is more likely to occur than the second category. If $\lambda_i^A < 0$, the first category is less likely to occur. If $\lambda_{ij}^{AB} = 0$ for all values of i and j, then variables A and B are independent. $\lambda_{ij}^{AB} > 0$ indicates the diagonal cells of agreement are more likely than the marginal table counts can account for, whereas $\lambda_{ij}^{AB} < 0$ indicates the diagonal cells of disagreement are more likely. Saturated models have no degrees of freedom and reproduce the observed cell frequencies exactly.

Unsaturated or restricted models do not include parameters for all possible effects. For example, an "independence model" constructed from Equation 3.1 contains no interaction effect between A and B. If $\lambda_{ij}^{AB} = 0$, then the log-probability of an observation in each cell can be adequately explained without the interaction parameter in the model.

$$\upsilon_{ij} = \overline{\upsilon}.. + \lambda_i^A + \lambda_j^B \tag{3.2}$$

In practice, if the interaction effect is minimal, a contingency table can be adequately described using an independence model. If the independence model adequately reproduces the cell counts, then it can be assumed that A and B were independent, meaning that the probability of occurrence in the i,jth cell is simply a product of the corresponding marginal probabilities. If the independence model does not adequately reproduce the cell counts, then the interaction effect is needed in the model. The degree to which the model adequately reproduces the observed cell counts is assessed in a goodness-of-fit test. The likelihood ratio statistic G^2 models the agreement between the observed cell counts and the modeled cell counts and is distributed as $\chi^2(\upsilon)$ where υ represents the model degrees of freedom. A small value indicates close fit while larger values indicate a misfitting model.

A "conditional equiprobability model" is a restricted model in which the interaction effects and one of the main effects are thought to be zero. Equation 3.2 can be restricted further by removing the main effect of the B variable. The conditional equiprobability model is shown in Equation 3.3.

$$\upsilon_{ij} = \overline{\upsilon}.. + \lambda_i^A \tag{3.3}$$

By excluding the main effect of B as in Equation 3.3, all levels of B are modeled as equally likely to occur, for each level of A.

An "equiprobability model" is the most restrictive model. Here, both the interaction and main effects are zero. An equiprobability model assumes that an observation is equally likely to occur in any cell. This model is shown in Equation 3.4.

$$v_{ij} = \overline{v}.. \qquad (3.4)$$

Like the independence model, the adequacy of conditional equiprobability models and equiprobability models can be assessed using goodness-of-fit procedures. The process of choosing a log-linear model to represent observed data involves eliminating superfluous effects resulting in a parsimonious model, yet ensuring that the model adequately explains the variations in the observed counts for each cell of the table (Kennedy, 1983). This overview used a 2 x 2 table for the purposes of simplicity. An advantage of log-linear models is that 3-way, 4-way, or larger contingency tables can be modeled and that the number of levels of a variable may be greater than two.

For the analysis of research question I - 1, each school/grade unit was represented in one cell of a cross-classification table of (method(i) x method(j) x grade(k)). This classification scheme prevents the same unit from appearing in more than one cell of the table. For example, all fourth grade units achieving OYG from both methods would appear as the counts in one particular cell, while all the fourth grade units achieving OYG from only one method would appear as the counts in a different cell. The four independent variables were Method AZ, Method A1, Method A2, and grade level. The dependent variable was the count of the number of schools in each cell of the table. The analysis for this research question did not use a four-way model AZ x A1 x A2 x Grade because it would have included the subordinate three-way interaction AZ x A1 x A2. The three-way interaction was not of interest because it would indicate that the pattern of interaction between two methods is different across levels of a third method and is not

useful in choosing methods. Further arguments are presented in Chapter 4 for not using the four-way model supported with results from the analyses.

Log-linear independence models were constructed to test for the presence or absence of interaction effects between method pairs and grade level. A hierarchical structure of effects was maintained by leaving lower-order effects in the models; that is, if an interaction effect AB was significant, then main effects A and B were also included regardless of their level of significance. More details of the specific analyses are provided in Chapter 4.

Analysis of Research Question I - 2

Research question I - 2: Do the OYG decisions of Methods AZ, A1, and A2 agree for each school/grade unit for Reading and Mathematics? To answer this question, Cohen's kappa (κ) and kappa max (κ_{max}) were used to analyze the agreement between the methods (Cohen 1960).

Cohen's κ is used to measure agreement between two ratings of the same object. κ is used for square two-way, cross-classification tables with two or more categorical levels. A value of 1 indicates perfect agreement, -1 perfect disagreement, and 0 that agreement is no better than chance. κ differs from a simple measure of the proportion of agreement by accounting for chance agreement. The equation for κ is

$$\kappa = \frac{p_O - p_C}{1 - p_C} \tag{3.5}$$

where p_O is the observed proportion of agreement (the sum of the observed cell proportions along diagonal cells of agreement), and p_C is the proportion of agreement

expected by chance (the sum of the expected proportions along diagonal cells of agreement). The expected proportion of agreement in each diagonal cell is the product of the corresponding marginal probabilities (under the hypothesis of independence).

The maximum value of 1.0 occurs when the off-diagonal cells of the cross-classification table are zero, and can only happen when the marginal distributions are identical to each other (i.e., when $p_{i.} = p_{.i}$ for every value of i). If this is not the case, then the maximum possible value of κ, κ_{max}, is less than 1.0. Cohen (1960) provides an equation to compute κ_{max} (Equation 3.6), where κ_{max} is the maximum value κ could achieve given the marginal values of the table. The equation for κ_{max} is

$$\kappa_{\max} = \frac{p_{OM} - p_C}{1 - p_C} \tag{3.6}$$

where p_{OM} is the maximum possible observed proportion of agreement. It is found by pairing the marginal values, selecting the smaller of each pair, and summing the values. Cohen (1960) suggests using the ratio of κ / κ_{max} as a measure of the amount of agreement given the maximum possible.

For the analysis of research question I - 2, OYG agreement was computed using κ and κ / κ_{max} for each pair of methods for each grade level to examine differences in the level of agreement from one grade to the next. These analyses were conducted separately for Reading and Mathematics.

Due to the large number of κ statistics computed in this analysis, the magnitudes were not tested for significance. The patterns of the agreement and the relative magnitudes between method pairs and grade levels were interpreted in a descriptive

manner. The κ values were used to indicate the absolute magnitude of the agreement (corrected for chance) between method pairs. The κ / κ_{max} values were used to indicate the agreement relative to the maximum possible.

Analysis of Research Question I - 3

Research question I - 3: Does the size of the school/grade unit affect the agreement of OYG between Methods AZ, A1, and A2 for Reading and Mathematics? The purpose of this research question is to determine whether method pairs may show less agreement in smaller unit sizes due to instability associated with small sample sizes. Comparing the agreement between method pairs grouped by the unit size will determine if sample size is a factor than needs to be considered while making a choice between the methods.

The analysis to answer this question was the same as question I - 2, except the school/grade units were grouped based on the size of each unit. To construct the groups the school/grade units were assigned to one of five size groupings based on the number of students in the unit (8-30, 31-60, 61-90, 91-120, and >120). For the purposes of comparing across grades, the same size groupings were used for all grade-levels. The agreement indicator κ and the ratio of κ / κ_{max} were computed for each pair of methods by size group within each grade level. As before, the κ values were not tested for statistical significance. Rather, they were examined and interpreted descriptively to highlight differences between the methods at different grade levels and different sizes.

Analysis of Research Question I - 4

Research question I - 4: Do the results differ among Methods AZ, A1, and A2 for computing Growth Rankings (the Star Rating) for Reading and Mathematics? The agreement indicator κ and the ratio of κ / κ_{max} were computed for each pair of methods at the grade level. κ and κ / κ_{max} were computed to assess agreement in Star Ratings between pairs of methods for each grade level.

The emphasis of this analysis was to determine similarities and differences between growth rankings computed using the three methods. If the rankings from one method were vastly different from the rankings of others, then the results may indicate method-dependency. Again, the κ values were not tested for statistical significance. Rather, they were interpreted descriptively to find differences between the methods at different grade levels.

Analysis of Category II:

The Effects of Inappropriately Correcting for RTM

The second set of research questions examined the impact of Arizona's correction for RTM on the OYG decision and Star Rating for each school/grade unit. Three research questions focused on differences and similarities between Arizona's method (Method AZ) and a variation of that method without a correction for RTM. Each research question and corresponding analysis procedure are listed separately.

Conditions to Examine the Effects of Correcting for RTM

To isolate the impact of correcting for RTM, the Arizona method was compared to an alternative with the same procedures in Method AZ except for the correction. The

alternative, Method $AZ_{NC.}$ was devised by calculating simple growth for each school/grade unit by subtracting the 1999 scaled score mean from the 1998 scaled score mean ($\bar{x}_{SS99} - \bar{x}_{SS98}$). This growth was then compared to the expected growth at the 50th percentile to make the OYG decision for each unit. A new Star Rating was then constructed using simple growth. The results of calculations from Methods AZ and AZ_{NC} were compared to answer the research questions in Category II.

The matched-student Stanford 9 achievement data were used for all of the questions in this category. The analyses for questions II-1 and II-3 began by computing the OYG decision separately for each school/grade unit using Methods AZ and AZ_{NC}. These analyses resulted in a dataset with two dichotomous indicators of OYG for every school/grade unit. For question II-2, Star Ratings were computed for each school/grade unit using Methods AZ and AZ_{NC}. The two Star Ratings and two OYG decisions for each school/grade unit were used for making comparisons between the methods. All analyses in this category were completed separately for Reading and Mathematics.

Analysis of Research Question II - 1

Research question II - 1: Do the OYG decisions of Methods AZ and AZ_{NC} agree for each school/grade unit for Reading and Mathematics? Due to the correction for RTM, the same school/grade unit may or may not achieve OYG between the two methods. To answer this question, κ and the ratio κ / κ_{max} were used to analyze the agreement between the two methods at each grade level. The κ values were not tested for statistical significance. Rather, the lack of agreement by grade level between Methods AZ and AZ_{NC} was used to describe the effects of the correction for RTM.

Analysis of Research Question II - 2

Research question II - 2: Do Methods AZ and AZ_{NC} give different Growth Rankings (the Star Rating) for Reading and Mathematics? The emphasis of this analysis was to determine the differences between growth rankings with and without the correction for RTM. For this question, κ and κ / κ_{max} were computed to assess the agreement in the Star Ratings between the two methods for each grade level. κ and κ / κ_{max} values were interpreted descriptively by examining the patterns of the agreement and the relative magnitudes across grade levels.

Analysis of Research Question II - 3

Research question II - 3: Does the initial scaled score mean for the school/grade unit affect the agreement for the OYG decision between Methods AZ and AZ_{NC} for Reading and Mathematics? The impact of correcting for RTM will be greater on units with initial means far above or below the overall mean for all units within a grade level.

To answer this research question, school/grade units were grouped based on initial scaled score status of each unit. For the groups the school/grade units were assigned quintiles based on their 1998 scaled score means. The agreement indicator κ and κ / κ_{max} were computed between the methods within each quintile for each grade level. The κ and κ / κ_{max} values were examined and interpreted descriptively for differences between the methods at different grade levels and mean scaled score groupings.

Analysis of Category III

The Independence of the Change Indicators

The third set of research questions examined the independence of the change indicators and initial status of the school/grade unit. The two research questions for this category addressed the degree to which growth, as measured in Methods AZ, AZ_{NC}, A1, A2, and A3, was uncorrelated with the initial status of the school/grade units. The first research question examined the amount of growth (simple, adjusted, or proportion OYG), the second research question examined the dichotomous OYG decision as an indicator of growth.

The matched-student Stanford 9 achievement data were used for both questions in this category. The analysis procedures for each research question are listed separately. The specific indicators (growth and initial status) computed for each method are listed in the following sections.

Analysis of Research Question III - 1

III - 1. What are the correlations between the amount of growth and initial status of the school/grade unit for Methods AZ, AZ_{NC}, A1, A2, and A3?

For Methods AZ and AZ_{NC}, the amount of growth was computed as the adjusted (Method AZ) or unadjusted (Method AZ_{NC}) scaled score growth mean from 1998 to 1999 for each school/grade unit. Initial status was measured as the mean 1998 scaled score.

For Method A1, the amount of growth was computed as the NCE growth mean from 1998 to 1999 of the students in each unit. Initial status was computed as the mean 1998 NCE score for each unit.

For Methods A2 and A3, the amount of growth was computed as the proportion of students achieving OYG within each school/grade unit. Initial status was computed as the median 1998 PR score (Method A2) or 1998 Stanine (Method A3) for each unit.

To answer question III-1, Pearson's product moment correlation coefficient r was computed between the amount of growth indicator and the initial status for each unit. Pearson's r is an indicator of linear relationship between a pair of variables y and x, given as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \qquad (3.7)$$

The coefficient ranges from –1 to 1, with the sign indicating the direction of the relationship and the magnitude indicating the strength.

The correlations were computed separately for each of the five methods within each grade level. The magnitudes of the correlations were used to examine the relationship between the amount of growth and initial status.

Analysis of Research Question III – 2

III-2. What are the correlations between the OYG decision and initial status of the school/grade unit for Methods AZ, AZ_{NC}, A1, A2, and A3?

For Methods AZ and AZ_{NC} initial status for each unit was computed as the 1998 scaled score mean. For Methods A1, A2, and A3 initial status for each unit was computed as the 1998 NCE score mean, the 1998 PR score median, or the 1998 stanine median respectively. The dichotomous growth indicator for these analyses was the OYG decision

for each school/grade unit. These were computed separately using Methods AZ, AZ_{NC}, A1, A2, and A3.

For question III-2, a point-biserial (r_{pb}) correlation was computed between the initial status indicator and the OYG decision for each school/grade unit. A point-biserial correlation (r_{pb}) is a special case of Pearson's product moment correlation used when one variable is continuous (y) and the other is dichotomous (x). The correlations were computed separately for each of the five methods within each grade level. The magnitudes of the r_{pb} values were used to examine the independence of the OYG decisions and initial status in each method.

Analysis of Category IV:

Examining the Effects of Accounting for Error

The emphasis of the fourth set of research questions was to examine the impact of accounting for error in Methods A1 and A2 on the OYG decision for each school/grade unit. Making adjustments to the passing criteria will increase the number of units achieving OYG, but the degree to which it will do so is unknown. The results from Method A3, using Stanines, were compared to the variations of Methods A1 and A2 to examine similarities and differences in the number of units achieving OYG.

Conditions to Examine the Effects of Accounting for Error

Method $A1_{AE}$ was devised to examine the impact of the criteria adjustments to account for errors on Method A1. The procedures in Method $A1_{AE}$ are identical to Method A1 except at the point of the OYG decision for each school/grade unit. In

Method $A1_{AE}$ the grade level standard error from Equation 2.13 is subtracted from the OYG criterion. The adjusted criterion is compared to the observed growth mean for each school/grade unit to make the OYG decision.

Method $A2_{AE}$ was constructed to isolate the effects of criteria adjustments in Method A2. The procedures in Method $A2_{AE}$ are the same as in Method A2 except for the OYG decision. In Method $A2_{AE}$ the grade level standard error from Equation 2.14 is subtracted from the OYG criterion of 50%. The adjusted criteria are used to make the OYG decisions.

The matched-student Stanford 9 achievement data were used for all three of the questions in this category. The OYG decision for each school/grade unit was computed using Methods A1, $A1_{AE}$, A2, $A2_{AE}$, and A3. These analyses resulted in a dataset with five dichotomous indicators of OYG for every school/grade unit. All analyses in this category were completed separately for Reading and Mathematics.

Analysis of Research Question IV – 1

IV-1. Do the OYG decisions of Methods A1 and $A1_{AE}$ agree for each school/grade unit for Reading and Mathematics? Methods A1 and $A1_{AE}$ may result in different OYG decisions for the same school/grade unit. To examine the differences, κ and the ratio of κ / κ_{max} were used to analyze the OYG agreement between the decisions at each grade level. As in the previous analyses, the κ values were not tested for statistical significance. Rather, the lack of agreement by grade level between Methods A1 and $A1_{AE}$ was used to describe the effects of accounting for error.

Analysis of Research Question IV – 2

IV-2. Do the OYG decisions of Methods A2 and $A2_{AE}$ agree for each school/grade unit for Reading and Mathematics? The analysis for this question was identical the previous research question. κ and the ratio of κ / κ_{max} were used to analyze the OYG agreement between decisions using Methods A2 and $A2_{AE}$ at each grade level. The lack of agreement by grade level was used to describe the effects of accounting for error.

Analysis of Research Question IV – 3

IV-3. Do the OYG decisions of Methods $A1_{AE}$, $A2_{AE}$, and A3 agree for each school/grade unit for Reading and Mathematics? As described earlier, the manner in which error is accounted for in Methods $A1_{AE}$ and $A2_{AE}$ is conceptually different from that of Method A3. To examine the similarities and differences in OYG assignment to units, the agreement indicator κ and the ratio of κ / κ_{max} were computed for each pairing of Methods $A1_{AE}$, $A2_{AE}$, and A3. The κ values were examined and the magnitudes interpreted descriptively to highlight differences among the methods at different grade levels.

Analysis of Category V

Minimum School/Grade Unit Size Criteria

The final category was used to determine the number of school/grade units that would be omitted from analyses if different minimum unit size criteria were employed. The reason for having a minimum unit size criteria is to provide statistically acceptable accuracy in estimating academic growth of each unit. The more students included in

estimating unit growth, the more stable the estimates. The practical problem is that the more students required, the lower the number of school/grade units that will have growth estimates available.

A modified version of the matched-student Stanford 9 achievement data were used for the single question in this category. All the matching and inclusion criteria for students remained the same as the other analyses, except that the minimum unit size requirement was varied from 5 to 25.

Analysis of Research Question V – 1

V-1. What proportion of school/grade units are excluded for different minimum size criteria? The analysis from this question was simply the overall proportion of units not included in analyses due to insufficient unit size when the criteria was varied from 5 to 25 students per unit. The proportions were plotted and interpreted descriptively.

CHAPTER 4

RESULTS AND DISCUSSION

The results are grouped by the five categories described in Chapters 2 and 3. For each section, the summary statistics from the analyses and discussions of the main findings are presented. Summary tables of the OYG and Star Ratings from all the methods in the five research question categories are presented.

The OYG and Star Ratings for Each Method

The OYG indicator and the Star Rating for each school/grade unit were computed using Methods AZ, A1, A2, A3, AZ_{NC}, $A1_{AE}$, and $A2_{AE}$. The proportions of schools achieving OYG by grade level using each of the methods are presented in Table 4.1 for Reading and Table 4.2 for Mathematics.

Table 4.1

Proportions of Schools Achieving OYG by Grade Level, Reading

Grade Level	AZ	A1	A2	A3	AZ_{NC}	$A1_{AE}$	$A2_{AE}$
3-4	.8694	.8631	.9020	.9860	.8802	.9456	.9487
4-5	.2610	.4977	.5645	.9712	.2929	.7633	.7678
5-6	.8517	.7933	.8494	.9910	.8112	.9258	.9393
6-7	.5430	.5781	.6914	.9727	.5234	.7656	.8359
7-8	.8343	.7259	.8133	.9880	.7771	.8283	.8976

Table 4.2

Proportions of Schools Achieving OYG by Grade Level, Mathematics

Grade Level	AZ	A1	A2	A3	AZ_{NC}	$A1_{AE}$	$A2_{AE}$
3-4	.7623	.7870	.8241	.9676	.7639	.9136	.9136
4-5	.6687	.6732	.7080	.9455	.6778	.8260	.8245
5-6	.8864	.8909	.9243	.9955	.9198	.9555	.9621
6-7	.4570	.4648	.5508	.9141	.5039	.5898	.6953
7-8	.4759	.5873	.6325	.9669	.4910	.6898	.7651

The data in Tables 4.1 and 4.2 suggest that the OYG decisions for school/grade units are not consistent across the different methods. For example, in the 4-5 grade level Reading, there is a large drop in the proportion of schools achieving OYG when Methods AZ or AZ_{NC} are used. The same drop is not as apparent for Mathematics for the same grade level. The methods that account for error (Methods A3, $A1_{AE}$, and $A2_{AE}$) show a consistently higher proportion of schools achieving OYG than the methods that do not. The differences between the methods are explained in greater detail from the results and discussions in the sections that follow.

For Methods AZ, A1, A2, A3, and AZ_{NC}, the cut-off values used to assign the Star Ratings to each school/grade unit are provided in Appendix B. The adjustments to account for error in Methods A1 and A2 have a constant impact on the magnitude of growth, so the Star Ratings for Method $A1_{AE}$ are the same as Method A1 and the Star Ratings for Method $A2_{AE}$ are the same as Method A2.

Results for Category I:

The Impact of the Method Choices in Arizona's Approach

The analyses from Category I were designed to examine the differences between Methods AZ, A1, and A2. Plots of the proportions of schools achieving OYG from each of the three methods are provided in Figures 4.1 and 4.2.

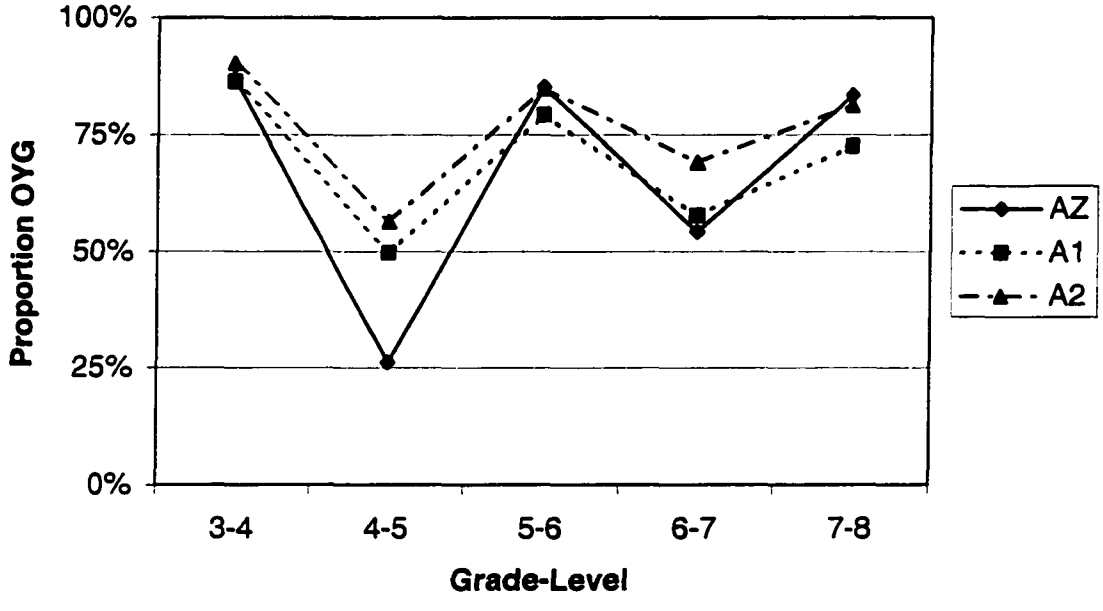

Figure 4.1. Proportions of schools achieving OYG by grade level using Methods AZ, A1, and A2 in Reading.

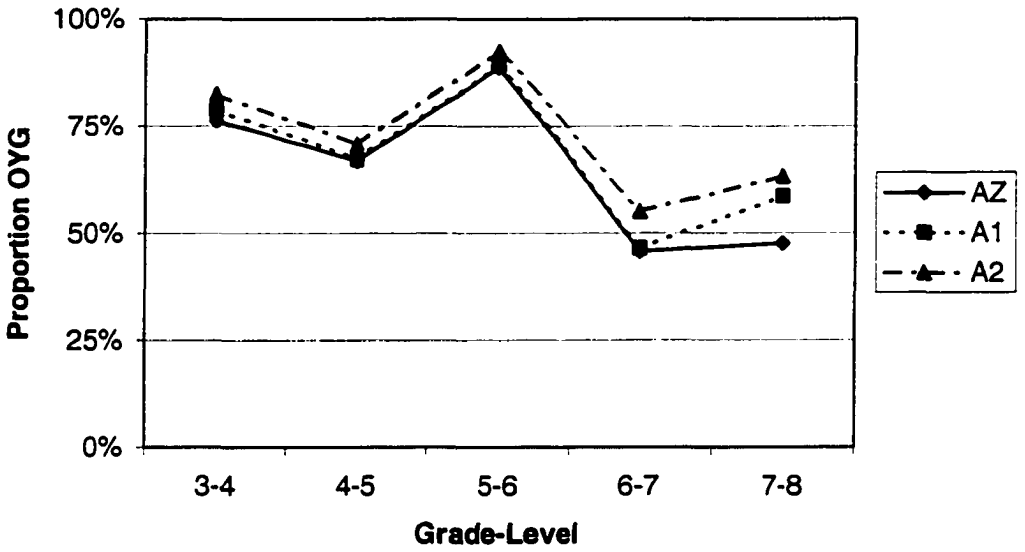

Figure 4.2. Proportions of schools achieving OYG by grade level using Methods AZ, A1, and A2 in Mathematics.

From the plots in Figures 4.1 and 4.2 the drop in proportion of schools achieving OYG from grades 4 to 5 for Method AZ can be seen for Reading but not for Mathematics. Method AZ appears to fluctuate more with respect to the other methods for Reading than for Mathematics. Method A2 yields consistently higher proportions than Method A1 for both Reading and Mathematics.

Results for Research Question 1 - 1

These analyses examined the differences among Methods AZ, A1, and A2 in the assignment of OYG to schools across grade levels. The log-linear analyses used four categorical variables. Outcome of computations using Methods AZ, A1, and A2 were coded as three dichotomous variables (1 or 0 depending on whether the school/grade unit achieved OYG using the given method). Grade level was the fourth variable coded with

five categories representing the grade level change of '3-4', '4-5', '5-6', '6-7', and '7-8'. The analysis for this research question did not use a four-way model AZ x A1 x A2 x Grade because it would have included the subordinate three-way interaction AZ x A1 x A2. As mentioned before, the three-way interaction was not of interest because it would indicate that the pattern of interaction between two methods is different across levels of a third method.

The three-way interaction of two competing methods crossed by grade level has a meaningful interpretation for this research. For example, if two methods assigned OYG to schools in a consistent manner across grade levels, then we can surmise that (a) neither of the methods has a specific grade level interaction or (b) they both have the same grade level interaction. The plot in Figure 4.1 shows that all three methods result in variations in the proportion of schools achieving OYG at each grade level. Methods A1 and A2 show a fairly consistent relationship with one other across grades suggesting they are functioning in a similar manner across grade levels. While Methods A1 and A2 appear to be assigning OYG to approximately the same proportion of schools, it is unknown if they are assigning OYG to the same schools. A log-linear analysis was used to answer this question. In the log-linear analyses, the three-way interaction of A1 x A2 x Grade was examined to determine if Methods A1 and A2 function consistently across grade levels. Figure 4.1 suggests that Method AZ does not have a consistent relationship with Methods A1 and A2 across grades, particularly at the '4-5' grade level change. The three-way interactions of AZ x A1 x Grade and AZ x A2 x Grade were examined to determine if

Method AZ assigned OYG to schools in a consistent manner as either Method A1 or Method A2.

Six log-linear analyses were conducted to examine the three-way interactions based on the following combinations of methods and grade: AZ x A1 x Grade (Reading and Mathematics), AZ x A2 x Grade: (Reading and Mathematics), and A1 x A2 x Grade: (Reading and Mathematics). Each of the six analyses began with saturated models. Log-linear analyses using an iterative proportional fitting procedure resulted in parameter estimates for each model as well as the test for fit (G^2) of the model to the observed data. Effects were removed hierarchically from each model, and the parameters and the fit of the new models were re-computed. For example, the three-way interaction effect was removed first and the model fit (G^2) re-assessed. If the new restricted model still fit the observed data, then two-way interaction effects were removed one at a time and the fit re-assessed. This procedure was followed until each analysis resulted in the most parsimonious model that still adequately described the observed data. To aid in following the results, the variables used in the six sets of analyses are listed in Table 4.3.

Table 4.3

Variables Included in Each of the Log-Linear Analyses

Analysis #	Variables	Area
1	AZ x A1 x Grade	
2	AZ x A2 x Grade	Reading
3	A1 x A2 x Grade	
4	AZ x A1 x Grade	
5	AZ x A2 x Grade	Math
6	A1 x A2 x Grade	

Log-linear analysis #1. Using the variables AZ, A1, and Grade for the Reading data, a saturated model was constructed based on AZ x A1 x Grade. To assess goodness-of-fit, the Likelihood Ratio G^2 was chosen over the Pearson χ^2 because the former can be linearly partitioned to assess the contribution of a particular effect to the fit of the model (Haberman, 1979). For a saturated log-linear model, $G^2(0) = 0$. It will reproduce the observed data exactly as there are as many parameters to estimate as df. The three-way interaction effect of AZ x A1 x Grade had four independent parameters to estimate (since grade has five levels and the effects are constrained to sum to 0) and thus four degrees of freedom. Upton (1978) suggests the examination of magnitudes for the standardized values of parameter estimates to assess the impact of an effect in the log-linear model. In Analysis #1 the standardized values for the three-way interaction ranged from –3.09 to 4.04. The large standardized parameter estimates for the three-way interaction suggested that they were needed to explain the observed data.

To assess the contribution of the three-way interaction in the model, the three-way interaction was removed (the parameter was constrained to zero) and the model re-estimated. For this model G^2 (4) = 59.10, and ($p < .001$). Since removing the three-way interaction significantly reduced the fit of the model, it was needed to explain the observed data. This means that the saturated model should be used to interpret the relationship between Methods AZ and A1 across grade levels. The resulting log-linear model for AZ x A1 x Grade (Reading) was

$$\upsilon_{ijk} = \overline{\upsilon}... + \lambda_i^{AZ} + \lambda_j^{A1} + \lambda_k^{Grade} + \lambda_{ij}^{AZxA1} + \lambda_{ik}^{AZxGrade} + \lambda_{jk}^{A1xGrade} + \lambda_{ijk}^{AZxA1xGrade}. \qquad (4.1)$$

The log-linear analysis produced estimates for each of the parameters in the model. Only the three-way interaction is discussed because it would not make sense to interpret lower-order effects when the presence of a significant higher-order effect suggests that they are not consistent.

To demonstrate how the parameter estimates can be used to reproduce cell probabilities, all the parameters needed to estimate the probability of the first cell (1,1,1) from Analysis #1 have been entered into Equation 4.1 resulting in

$$\upsilon_{111} = -3.838 - .145 - .372 - .082 + 1.031 - .437 - .035 + .412.$$

The log-probability of an occurrence in cell (1,1,1) is υ_{111} = -3.47. This relates to a probability of .0312. The actual proportion of observations in this cell was .0321. Barring rounding errors, the saturated model will reproduce the observed data.

Table 4.4

Three-Way Interaction Parameter Estimates for the AZ x A1 x Grade (Reading) Analysis

		A1		
		0	1	
AZ	0	.412	-.412	Grade 3-4
	1	-.412	.412	
AZ	0	.339	-.339	Grade 4-5
	1	-.339	.339	
AZ	0	-.013	.013	Grade 5-6
	1	.013	-.013	
AZ	0	-.263	.263	Grade 6-7
	1	.263	-.263	
AZ	0	-.475	.475	Grade 7-8
	1	.475	-.475	

The un-standardized parameters in Table 4.4 show that Methods AZ and A1 tended to agree more than average on the OYG decision for schools in Reading in grade levels 3-4 and 4-5, (note the positive values along the diagonal of agreement). Methods AZ and A1 tended to disagree in grade levels 6-7 and 7-8, (note the negative values along the diagonal of agreement). This reversal of patterns suggests that Methods AZ and A1 do not assign OYG to the same schools in a consistent manner across grade levels.

Log-linear analysis #2. Analyses #2 (AZ x A2 x Grade for Reading) resulted in a significant three-way interaction effect just as in Analysis #1, so the saturated model was

retained as the best fitting model. Because the procedures for each analysis were the same as Analysis #1, the individual steps of the analyses are not detailed. The log-linear models and the parameter estimates for the three-way interaction from the analysis are provided. For Analysis #2 the resulting log-linear model was

$$\upsilon_{ijk} = \overline{\upsilon}... + \lambda_i^{AZ} + \lambda_j^{A2} + \lambda_k^{Grade} + \lambda_{ij}^{AZxA2} + \lambda_{ik}^{AZxGrade} + \lambda_{jk}^{A2xGrade} + \lambda_{ijk}^{AZxA2xGrade}. \qquad (4.2)$$

The three-way interaction had a fit of $G^2 = 27.48$ and ($p < .001$). Table 4.5 shows the parameter estimates for the three-way interaction.

Table 4.5

Three-Way Interaction Parameter Estimates for the AZ x A2 x Grade (Reading) Analysis

		A2		
		0	1	
AZ	0	.199	-.199	Grade 3-4
	1	-.199	.199	
AZ	0	.288	-.288	Grade 4-5
	1	-.288	.288	
AZ	0	-.058	.058	Grade 5-6
	1	.058	-.058	
AZ	0	-.100	.100	Grade 6-7
	1	.100	-.100	
AZ	0	-.329	.329	Grade 7-8
	1	.329	-.329	

The parameters in Table 4.5 show that Methods AZ and A2 tended to agree more on the OYG decision for schools in Reading in grade levels 3-4 and 4-5 but tended to disagree in grade levels 6-7 and 7-8. Methods AZ and A2 do not assign OYG for Reading to the same schools in a consistent manner across the early to later grade levels.

Log-linear analysis #3. Using the variables A1, A2, and Grade for the Reading data, a total of five hierarchical models were examined for fit of the observed data. Table 4.6 shows a summary of the models tested for Analysis #3 ("X" indicates that a parameter was included in the model and estimated).

Table 4.6

Models for the A1 x A2 x Grade (Reading) Analysis

Model	Effect							df	G^2	p(G^2)
	λ_i^{A1}	λ_j^{A2}	λ_k^{Grade}	λ_{ij}^{A1xA2}	$\lambda_{ik}^{A1xGrade}$	$\lambda_{jk}^{A2xGrade}$	$\lambda_{ijk}^{A1xA2xGrade}$			
1	X	X	X	X	X	X	X	0	0.00	1.000
2	X	X	X	X	X	X		4	6.87	.140
3	X	X	X	X	X			8	45.50	< .001
4	X	X	X	X		X		8	55.00	< .001
5	X	X	X		X	X		5	1096.00	< .001

The analysis started with a saturated model (Model 1). The three-way interaction was constrained in Model 2 resulting in G^2 (4) = 6.87 and (p= .14), suggesting that the three-way interaction was not needed in the model. The two-way interactions were then removed from the model one at a time (Models 3, 4, and 5). The significant decrease in goodness-of-fit of these subsequent models lead to the retention of all the two-way

interactions. For the A1 x A2 x Grade (Reading) analysis, the following log-linear model was chosen

$$\upsilon_{ijk} = \overline{\upsilon}... + \lambda_i^{A1} + \lambda_j^{A2} + \lambda_k^{Grade} + \lambda_{ij}^{A1xA2} + \lambda_{ik}^{A1xGrade} + \lambda_{jk}^{A2xGrade}. \quad (4.3)$$

While the absence of the three-way interaction is the more important feature of this analysis, the model does have significant two-way interactions. Tables 4.7 – 4.9 show the un-standardized parameter estimates for the two-way interactions.

Table 4.7

Parameter Estimates for the A1 x A2 Interaction in the A1 x A2 x Grade (Reading) Analysis

		A2	
		0	1
A1	0	1.067	-1.067
	1	-1.067	1.067

The parameters in Table 4.7 show that Methods A1 and A2 had a greater than chance tendency to assign OYG to the same school (collapsed across grade levels).

Table 4.8

Parameter Estimates for the A1 x Grade Interaction in the A1 x A2 x Grade (Reading) Analysis

		Grade				
		3-4	4-5	5-6	6-7	7-8
A1	0	-.355	.203	-.164	.275	.041
	1	.355	-.203	.164	-.275	-.041

Table 4.9

Parameter Estimates for the A2 x Grade Interaction in the A1 x A2 x Grade (Reading) Analysis

		Grade				
		3-4	4-5	5-6	6-7	7-8
A2	0	-.184	.360	-.088	.032	-.120
	1	.184	-.360	.088	-.032	.120

The parameters from Tables 4.8 and 4.9 indicate that both Methods A1 and A2 did not assign an equal proportion of OYG to schools across grade levels. The interpretation of all three of the two-way interactions is somewhat secondary to the fact that the three-way interaction of A1 x A2 x Grade was not statistically significant for Reading.

Log-linear analyses #4, and #5. Analyses #4 and #5 each resulted in significant three-way interaction effects just as in Analysis #1, so the saturated model was retained as the best fitting model in each case. The procedures for each analysis were the same as Analysis #1, the individual steps of the analyses are not detailed.

For analysis #4 (AZ x A1 x Grade for Mathematics) the resulting log-linear model was Equation 4.1. The three-way interaction had a fit of G^2 (4) = 19.22 and (p = .001). Table 4.10 shows the parameter estimates for the three-way interaction.

Table 4.10

Three-Way Interaction Parameter Estimates for the AZ x A1 x Grade (Mathematics) Analysis

		A1		
		0	1	
AZ	0	-.109	.109	Grade 3-4
	1	.109	-.109	
AZ	0	.091	-.091	Grade 4-5
	1	-.091	.091	
AZ	0	.033	-.033	Grade 5-6
	1	-.033	.033	
AZ	0	.224	-.224	Grade 6-7
	1	-.224	.224	
AZ	0	-.239	.239	Grade 7-8
	1	.239	-.239	

The parameter estimates in Table 4.10 show that Methods AZ and A1 tended to agree more on the OYG (Mathematics) decision for schools in grade level 6-7 but tended to disagree in grade level 7-8. Methods AZ and A1 do not assign OYG for Mathematics to the same schools in a consistent manner across the later grade levels.

For Analysis #5 (AZ x A2 x Grade for Mathematics) the resulting log-linear model was Equation 4.2. The three-way interaction had a fit of G^2 (4) = 25.84 and ($p < .001$). Table 4.11 shows the parameter estimates for the three-way interaction.

Table 4.11

Three-Way Interaction Parameter Estimates for the AZ x A2 x Grade (Mathematics) Analysis

		A2		
		0	1	
AZ	0	-.236	.236	Grade 3-4
	1	.236	-.236	
AZ	0	.015	-.015	Grade 4-5
	1	-.015	.015	
AZ	0	.070	-.070	Grade 5-6
	1	-.070	.070	
AZ	0	.375	-.375	Grade 6-7
	1	-.375	.375	
AZ	0	-.224	.224	Grade 7-8
	1	.224	-.224	

Table 4.11 shows that Methods AZ and A2 tended to agree more on the OYG (Mathematics) decision for schools in grade level 6-7 but tended to disagree in grade levels 3-4 and 7-8. Once again, Methods AZ and A2 do not assign OYG for Mathematics to the same schools in a consistent manner across grade levels.

Log-linear analysis #6. Using the variables A1, A2, and Grade for the Mathematics data, five hierarchical models were examined for fit of the observed data.

Table 4.12 shows a summary of the models tested for Analysis #6 ("X" indicates that a parameter was included in the model and estimated).

Table 4.12

Models for the A1 x A2 x Grade (Mathematics) Analysis

Model	Effect							df	G^2	p(G^2)
	λ_i^{A1}	λ_j^{A2}	λ_k^{Grade}	λ_{ij}^{A1xA2}	$\lambda_{ik}^{A1xGrade}$	$\lambda_{jk}^{A2xGrade}$	$\lambda_{ijk}^{A1xA2xGrade}$			
1	X	X	X	X	X	X	X	0	0.00	1.000
2	X	X	X	X	X	X		4	3.27	.510
3	X	X	X	X	X			8	25.50	< .001
4	X	X	X	X		X		8	36.40	< .001
5	X	X	X		X	X		5	1322.00	< .001

The analysis proceeded in the same manner as Analysis #3, resulting in the removal of the three-way interaction effect but retention of the two-way interactions. For Analysis #6, the log-linear model chosen was Equation 4.3. The model fit was G^2 (4) = 3.27 and (p = .510). Tables 4.13 – 4.15 show the un-standardized parameter estimates for the two-way interactions of this model.

Table 4.13

Parameter Estimates for the A1 x A2 Interaction in the A1 x A2 x Grade (Mathematics) Analysis

		A2	
		0	1
A1	0	1.178	-1.178
	1	-1.178	1.178

The parameters in Table 4.13 show that Methods A1 and A2 had a greater than chance tendency to assign OYG in Mathematics to the same school (collapsed across grade levels).

Table 4.14

Parameter Estimates for the A1 x Grade Interaction in the A1 x A2 x Grade (Mathematics) Analysis

		Grade				
		3-4	4-5	5-6	6-7	7-8
A1	0	-.155	-.034	-.337	.416	.110
	1	.155	.034	.337	-.416	-.110

Table 4.15

Parameter Estimates for the A2 x Grade Interaction in the A1 x A2 x Grade (Mathematics) Analysis

		Grade				
		3-4	4-5	5-6	6-7	7-8
A2	0	-.077	.152	-.408	.128	.205
	1	.077	-.152	.408	-.128	-.205

The results from the six log-linear analyses demonstrated that a three-way interaction existed for the four analyses involving Method AZ, and did not exist between Methods A1 and A2 across grade levels. Table 4.16 shows a summary of the six analyses.

Table 4.16

Summary of the Significant Three-Way Interactions in the Six Log-Linear Analyses

Analysis #	Variables	Area	Sig. 3-way Interaction
1	AZ x A1 x Grade		Yes
2	AZ x A2 x Grade	Reading	Yes
3	A1 x A2 x Grade		No
4	AZ x A1 x Grade		Yes
5	AZ x A2 x Grade	Math	Yes
6	A1 x A2 x Grade		No

These six analyses suggest that Methods A1 and A2 perform in a consistent manner across grade level. While both reveal differences in the proportions of schools achieving OYG across grade levels, they both tend to reveal the same differences. Method AZ does not show the same pattern of proportions across grade levels and is not consistent with Methods A1 and A2.

Results for Research Question I - 2

These analyses examined the agreement between Methods AZ, A1, and A2 in the assignment of OYG to school/grade units. The results of κ and κ / κ_{max} analyses on the agreement of the OYG decisions for school/grade units using Methods AZ, A1, and A2 are presented in Table 4.17 and Figures 4.3 - 4.4 for Reading and Table 4.20 and Figures 4.5 – 4.6 for Mathematics. Rather than compute the significance of the κ values, the relative magnitudes were compared across different analyses rather than strictly interpreting significant results. The κ values were used to express the amount of

agreement between the methods being compared and the κ / κ_{max} values were used to express the amount of agreement given the maximum possible.

Table 4.17

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between each Pair of Methods AZ, A1, and A2, by Grades, Reading

Grade Level	AZ and A1			AZ and A2			A1 and A2		
	κ	κ_{max}	κ / κ_{max}	K	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}
3-4	.852	.973	.876	.655	.839	.781	.693	.813	.853
4-5	.519	.526	.988	.417	.428	.973	.715	.867	.825
5-6	.663	.801	.828	.567	.991	.572	.672	.810	.831
6-7	.644	.929	.693	.548	.693	.790	.726	.759	.956
7-8	.396	.689	.574	.388	.927	.419	.689	.756	.911

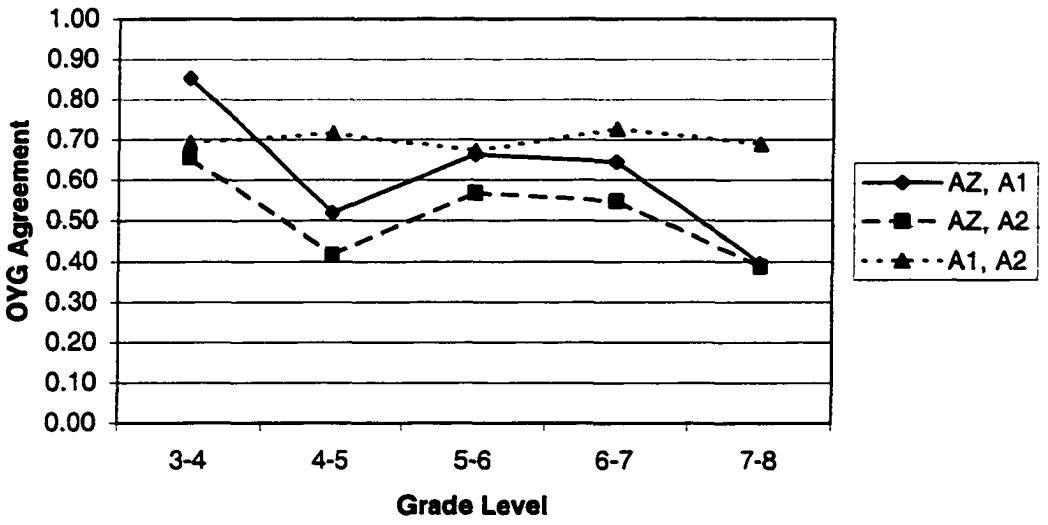

Figure 4.3. Plots of κ of the OYG decision between each pair of Methods AZ, A1, and A2, by grades in Reading.

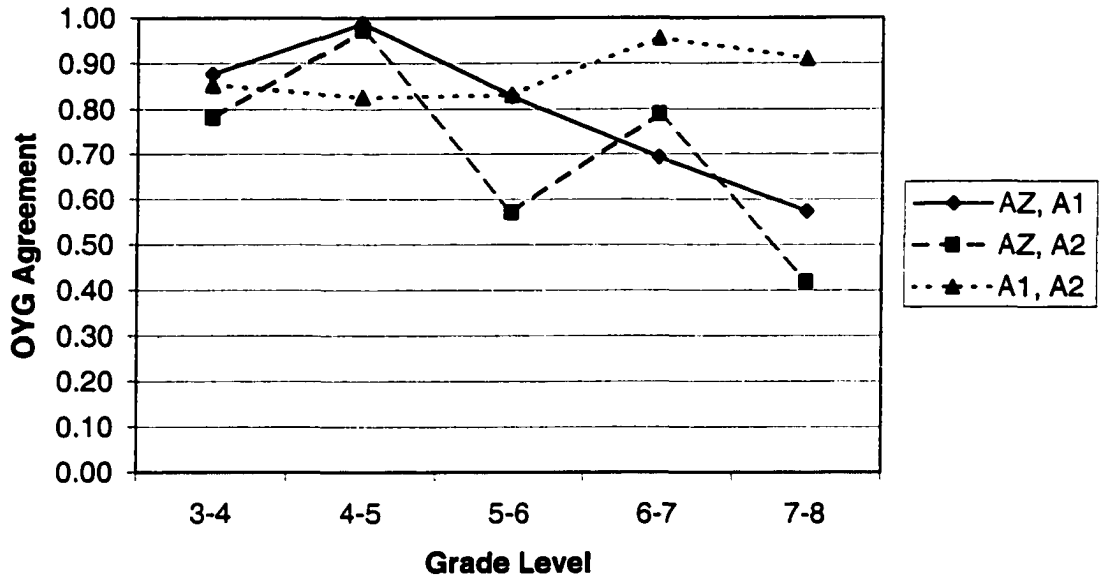

Figure 4.4. Plots of κ / κ_{max} of the OYG decision between each pair of Methods AZ, A1, and A2, by grades in Reading.

Table 4.17 and the plots in Figures 4.3 – 4.4 suggest that generally the agreement between Methods A1 and A2 is more consistent across grade levels than the agreement for Methods AZ and A1 or AZ and A2. While there is a decrease in the amount of agreement for Method AZ with A1 and A2 in grades 4-5, examining the κ / κ_{max} indicates that the agreement is about as large as it can be. This is because the proportion of schools classified as passing differs so greatly between methods for grades 4-5. In grades 7-8 there is a decrease in the amount of agreement between Method AZ with A1 and A2, and the κ / κ_{max} coefficients are lower indicating that the agreements are not as large as they could be given the maximum possible.

To aid in the interpretation of the κ / κ_{max} indicators, contingency tables showing the OYG agreement between Methods AZ and A1 are provided in Table 4.18 for the 4-5 grade level and Table 4.19 for the 7-8 grade level.

Table 4.18

OYG Agreement between Methods AZ and A1 Grade 4-5, Reading

		Method A1		
Method AZ		0	1	Total
0	Count	330	157	487
1	Count	1	171	172
Total	Count	331	328	659

Table 4.19

OYG Agreement between Methods AZ and A1 Grade 7-8, Reading

		Method A1		
Method AZ		0	1	Total
0	Count	38	17	55
1	Count	53	224	277
Total	Count	91	241	332

The counts in Table 4.18 show that there was only one school in the 4-5 grade level for which Method A1 did not assign OYG but Method AZ did. The marginal counts from this table show discrepancies between the numbers of schools each method identified as achieving OYG. Because the marginal counts are different from each other,

it is not possible for the methods to agree on every decision. The κ value for this table was .519, which is not a relatively large amount of agreement. The largest it could have been was .526. Therefore, the ratio of κ / κ_{max} of .988 showed that Methods AZ and A1 agreed almost as much as they could have agreed once chance agreement was accounted for.

The marginal counts in Table 4.19 show that in the 7-8 grade level, Methods AZ and A1 assigned OYG to approximately the same number schools (at least more similar than for grade level 4-5). As a result, it is possible for the methods to have better agreement in this grade level. The κ value for Table 4.19 was .396 and the largest it could have been was .698. Therefore, the ratio of κ / κ_{max} of .574 showed that Methods AZ and A1 did not agree as much as they could have.

Table 4.20

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between each Pair of Methods AZ, A1, and A2, by Grades, Mathematics

Grade Level	AZ and A1			AZ and A2			A1 and A2		
	κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}
3-4	.629	.929	.677	.523	.813	.643	.754	.882	.855
4-5	.764	.989	.772	.718	.908	.791	.806	.919	.877
5-6	.640	.977	.655	.599	.780	.768	.669	.802	.835
6-7	.827	.984	.840	.783	.814	.962	.798	.829	.963
7-8	.564	.779	.724	.536	.691	.776	.766	.905	.846

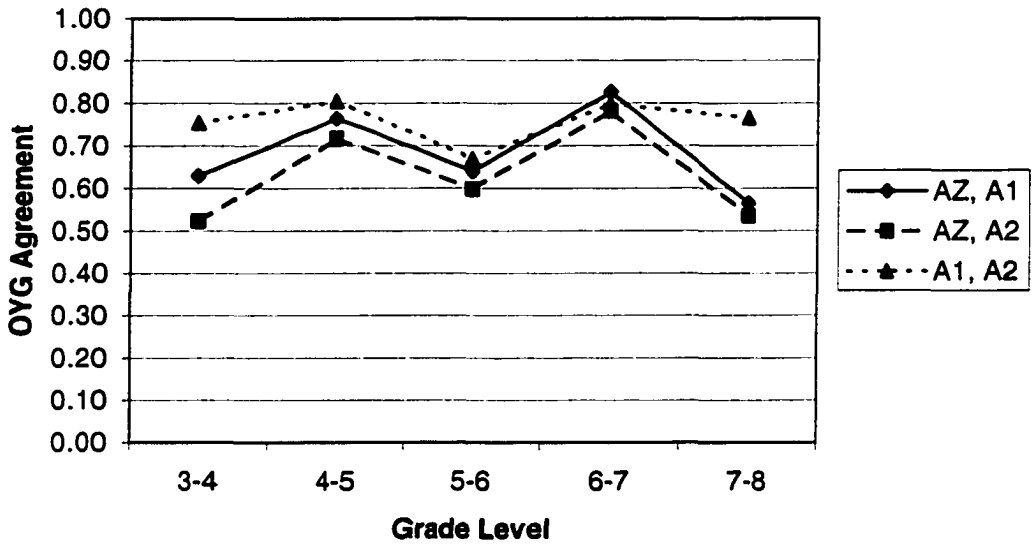

Figure 4.5. Plots of κ of the OYG decision between each pair of Methods AZ, A1, and A2, by grades in Mathematics.

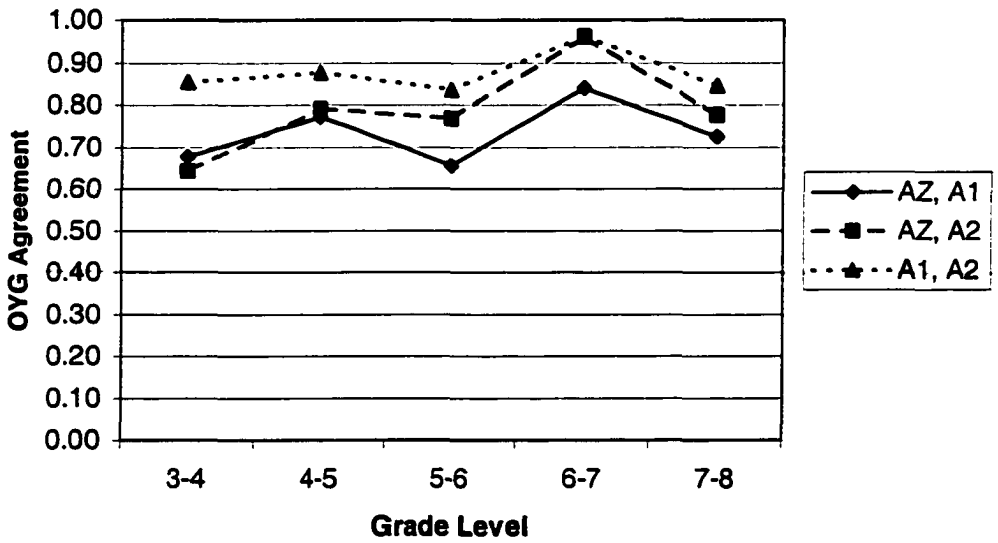

Figure 4.6. Plots of κ / κ_{max} of the OYG decision between each pair of Methods AZ, A1, and A2, by grades in Mathematics.

The results in Table 4.20 and Figures 4.5 – 4.6 suggest that the amounts of agreement between all the method pairs are not the same across grade levels. The lowest κ value is .523 between AZ and A2 at grade level 3-4. The κ and κ / κ_{max} values are higher for Methods A1 and A2 than for AZ with A1 or A2, indicating that Methods A1 and A2 are more likely to classify the same school/grade unit as passing, while Method AZ is less consistent.

Results for Research Question 1 - 3

The results of the κ and κ / κ_{max} analyses on the agreement of the OYG decisions by the size of the school/grade units using Methods AZ, A1, and A2 are presented in Table 4.21 and Figures 4.7 – 4.8 for Reading and Table 4.22 and Figures 4.9 – 4.10 for Mathematics.

It was expected that there would be lower agreement between Methods A1 and A2 when the school/grade unit size was smaller. The data in Tables 4.21 – 4.22 and Figures 4.7 – 4.10 do not support this expectation. There is pattern with AZ with A1 and AZ with A2 in Reading in which there is better agreement for units with fewer numbers of students than those with larger numbers. This pattern is not consistent in Mathematics for the same methods. These inconsistent findings seem to suggest that the number of students in a unit is not a critical factor in the performance of one method over another.

Table 4.21

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between each Pair of Methods AZ, A1, and A2, by Grades by Unit Size Groups, Reading

Grade Level	Size Group	Units / Group	AZ and A1			AZ and A2			A1 and A2		
			κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}
3-4	8-30	152	.930	1.00	.930	.714	.828	.862	.714	.829	.862
	31-60	241	.778	1.00	.778	.599	.810	.739	.683	.809	.844
	61-90	173	.751	.834	.900	.544	.949	.573	.613	.785	.781
	91-120	60	1.00	1.00	1.00	.792	.792	1.00	.792	.792	1.00
	>120	17	1.00	1.00	1.00	.541	.809	.668	.632	.887	.712
4-5	8-30	143	.523	.549	.952	.497	.549	.904	.675	1.00	.675
	31-60	239	.524	.524	1.00	.357	.357	1.00	.692	.759	.912
	61-90	183	.523	.523	1.00	.451	.451	1.00	.749	.902	.831
	91-120	67	.576	.576	1.00	.480	.480	1.00	.753	.876	.859
	>120	27	.138	.138	1.00	.104	.104	1.00	.702	.851	.825
5-6	8-30	90	.699	1.00	.699	.615	.872	.706	.744	.872	.853
	31-60	147	.749	.875	.857	.603	.851	.708	.686	.731	.939
	61-90	143	.565	.565	1.00	.472	.670	.705	.582	.877	.664
	91-120	44	.495	.495	1.00	.535	.845	.633	.621	.621	1.00
	>120	21	.447	1.00	.447	.447	1.00	.447	1.00	1.00	1.00
6-7	8-30	79	.688	1.00	.688	.604	.703	.859	.654	.703	.930
	31-60	48	.576	.765	.896	.610	.610	1.00	.716	.829	.863
	61-90	48	.652	.913	.714	.511	.777	.657	.862	.862	1.00
	91-120	25	.590	.918	.643	.233	.540	.432	.478	.478	1.00
	>120	56	.544	.876	.622	.519	.694	.748	.811	.811	1.00
7-8	8-30	72	.661	.952	.695	.413	.543	.760	.506	.506	1.00
	31-60	58	.404	.744	.542	.469	.952	.493	.622	.790	.787
	61-90	51	.254	.751	.338	.314	.842	.373	.717	.906	.792
	91-120	20	.432	.886	.487	.474	.737	.643	.634	.634	1.00
	>120	131	.307	.450	.682	.334	.637	.525	.767	.767	1.00

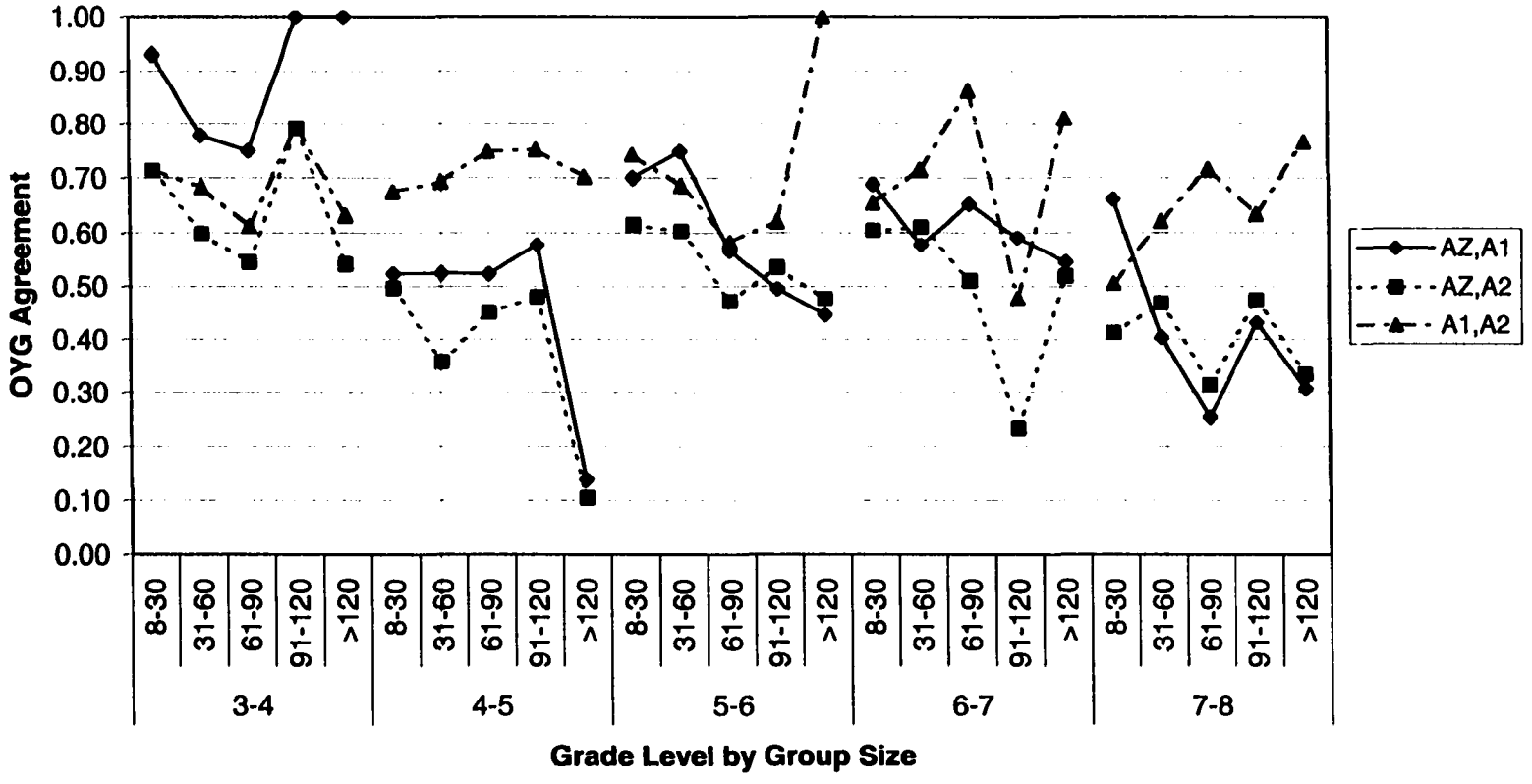

Figure 4.7. Plot of κ values of the OYG decision between each pair of Methods AZ, A1, and A2 by grade by unit size groups in Reading.

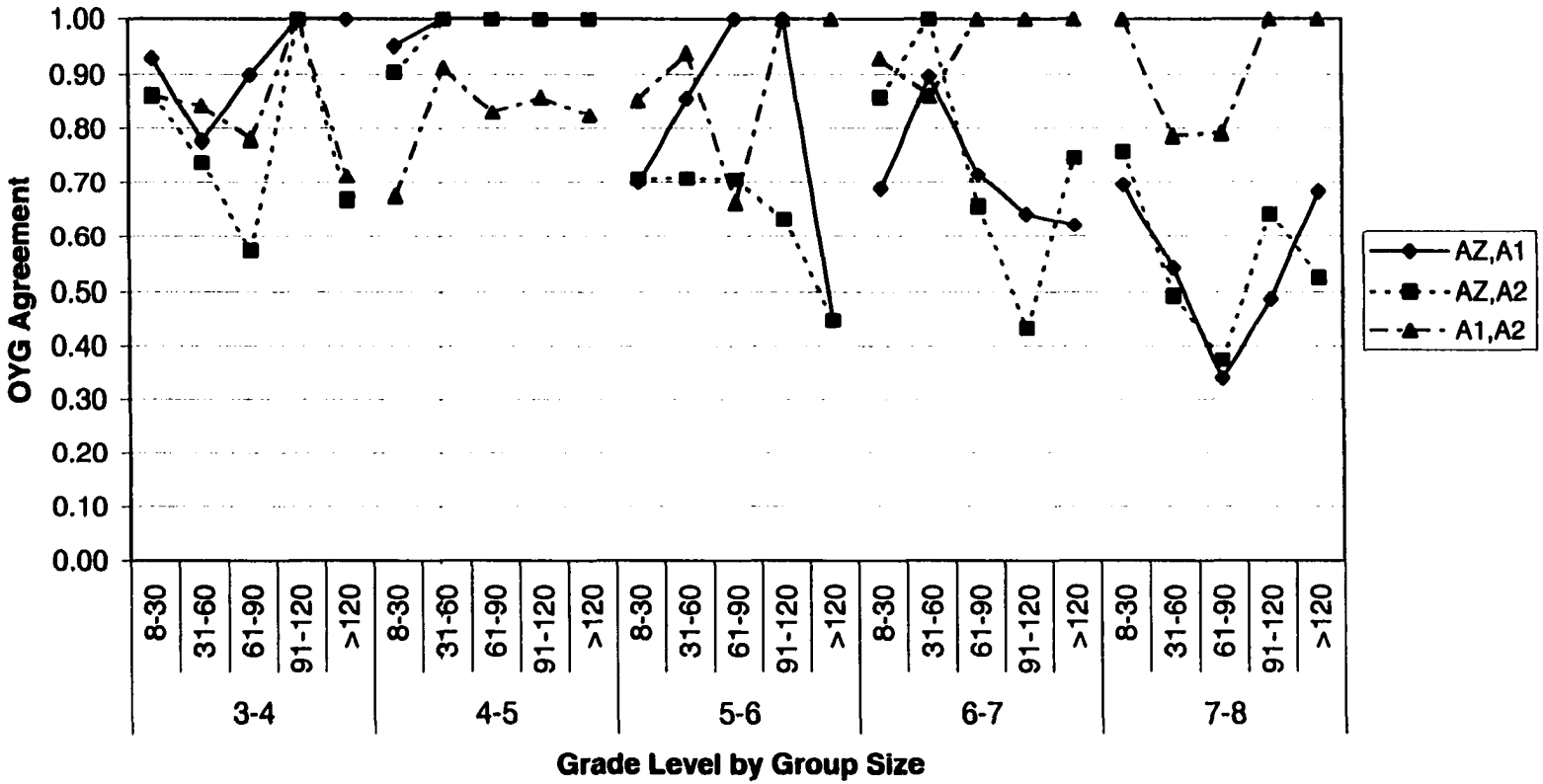

Figure 4.8. Plot of κ / κ_{max} values of the OYG decision between each pair of Methods AZ, A1, and A2 by grade by unit size groups in Reading.

Table 4.22

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between each Pair of Methods AZ, A1, and A2, by Grades by Unit Size Groups, Mathematics

Grade Level	Size Group	Units / Group	AZ and A1			AZ and A2			A1 and A2		
			κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}
3-4	8-30	147	.747	.781	.957	.506	.647	.782	.650	.856	.759
	31-60	241	.621	.800	.777	.541	.725	.747	.762	.921	.828
	61-90	172	.590	.795	.742	.602	.943	.638	.817	.850	.961
	91-120	64	.430	.857	.501	.170	.924	.184	.797	.932	.855
	>120	24	.059	.059	1.00	.043	1.00	.043	.647	.647	1.00
4-5	8-30	135	.844	.838	.900	.684	.873	.782	.774	.935	.827
	31-60	240	.722	.930	.775	.692	.868	.797	.793	.937	.846
	61-90	189	.848	.958	.884	.783	.957	.819	.859	.915	.938
	91-120	65	.594	.675	.880	.767	.860	.892	.728	.805	.904
	>120	32	.389	.796	.488	.459	.676	.680	.871	.871	1.00
5-6	8-30	94	.685	.895	.765	.598	.817	.731	.683	.921	.742
	31-60	145	.638	.819	.779	.671	.765	.877	.718	.944	.761
	61-90	144	.480	.851	.564	.274	.609	.450	.490	.490	1.00
	91-120	44	.845	.845	1.00	1.00	1.00	1.00	.845	.845	1.00
	>120	22	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6-7	8-30	77	.728	.891	.817	.720	.720	1.00	.765	.824	.929
	31-60	52	.884	.961	.919	.841	.841	1.00	.804	.804	1.00
	61-90	45	1.00	1.00	1.00	.897	.897	1.00	.897	.897	1.00
	91-120	23	.862	.862	1.00	.777	.777	1.00	.650	.650	1.00
	>120	59	.722	1.00	.722	.664	.798	.831	.731	.798	.916
7-8	8-30	72	.540	.655	.824	.597	.655	.912	.731	1.00	.731
	31-60	58	.430	.631	.681	.399	.599	.665	.811	.962	.843
	61-90	51	.720	.720	1.00	.615	.615	1.00	.805	.883	.911
	91-120	23	.819	.819	1.00	.582	.582	1.00	.743	.743	1.00
	>120	128	.511	.921	.555	.507	.793	.639	.740	.870	.850

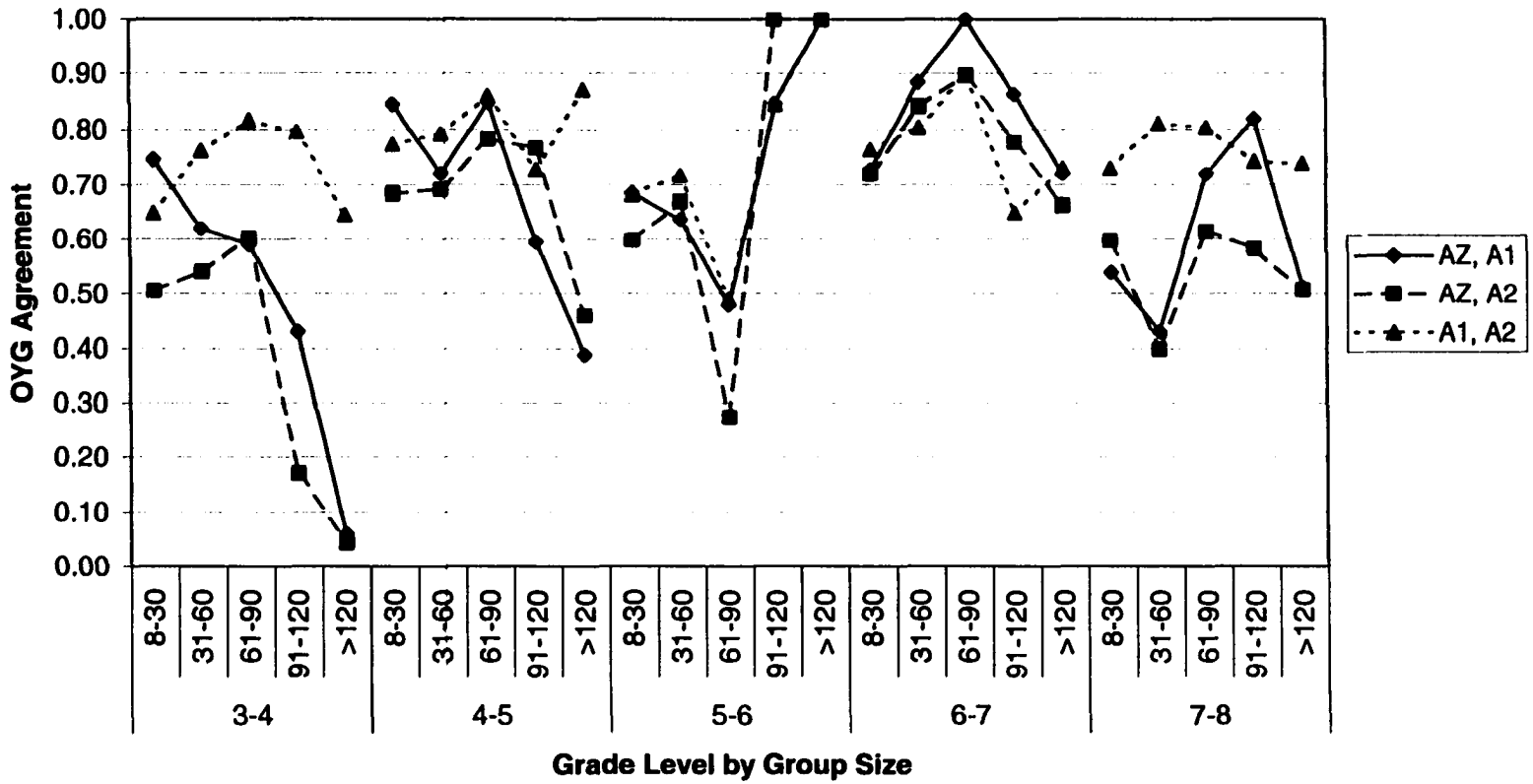

Figure 4.9. Plot of κ values of the OYG decision between each pair of Methods AZ, A1, and A2 by grade by unit size groups in Mathematics.

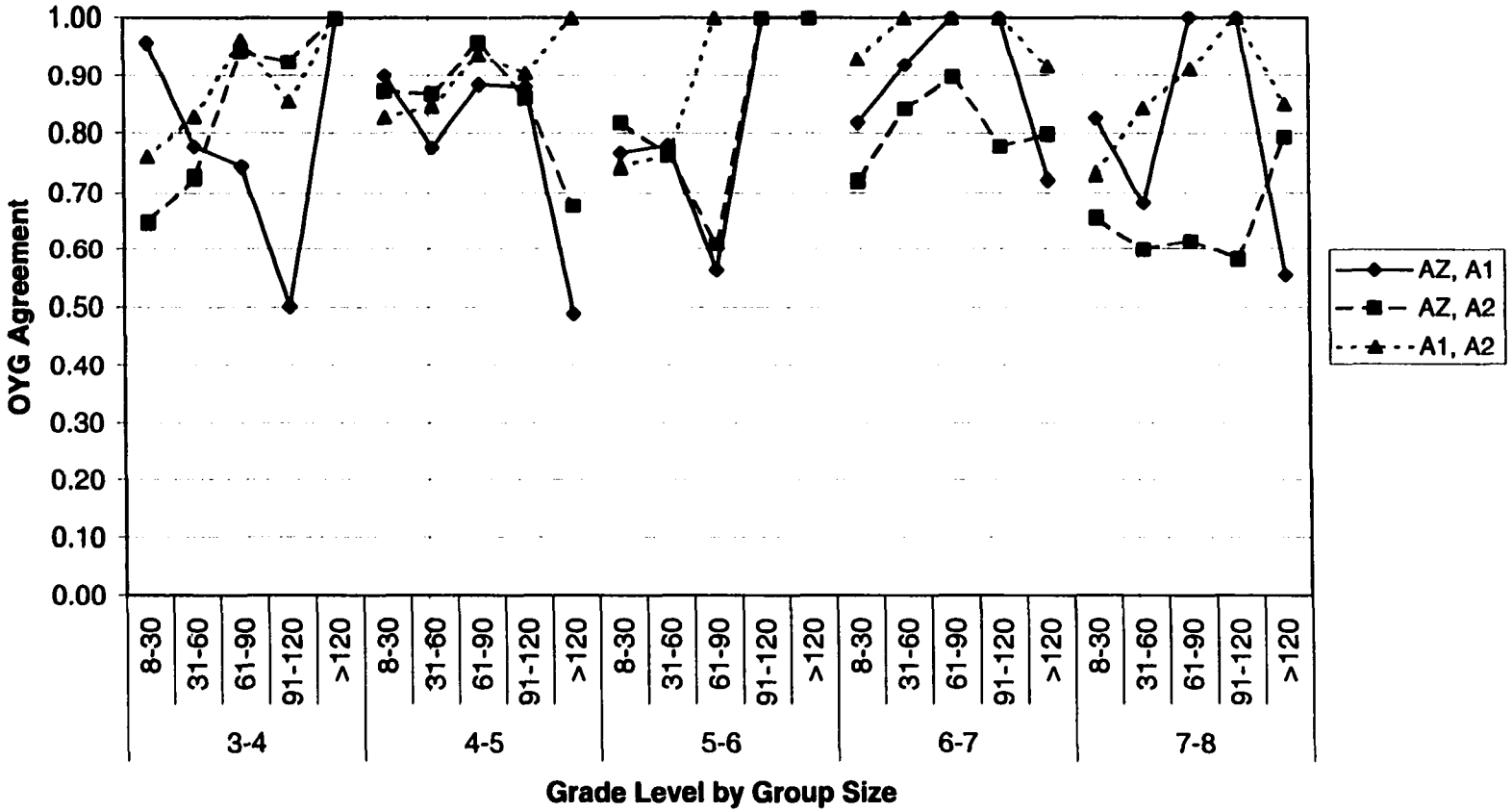

Figure 4.10. Plot of κ / κ_{max} values of the OYG decision between each pair of Methods AZ, A1, and A2 by grade by unit size groups in Mathematics.

Results for Research Question 1 - 4

The results of the κ and κ / κ_{max} analyses on the agreement of the Star Ratings for school/grade units using Methods AZ, A1, and A2 are presented in the Table 4.23 and Figures 4.11 – 4.12 for Reading and Table 4.24 and Figures 4.13 – 4.14 for Mathematics.

Table 4.23

κ, κ_{max}, and κ / κ_{max} of the Star Ratings between each Pair of Methods AZ, A1, and A2, by Grades, Reading

Grade Level	AZ and A1			AZ and A2			A1 and A2		
	κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}
3-4	.568	1.00	.568	.417	.996	.418	.487	.996	.489
4-5	.539	1.00	.539	.385	.996	.387	.495	.996	.497
5-6	.520	1.00	.520	.360	.997	.361	.475	.997	.476
6-7	.321	1.00	.321	.287	.990	.290	.526	.990	.532
7-8	.292	1.00	.292	.239	.989	.242	.563	.989	.570

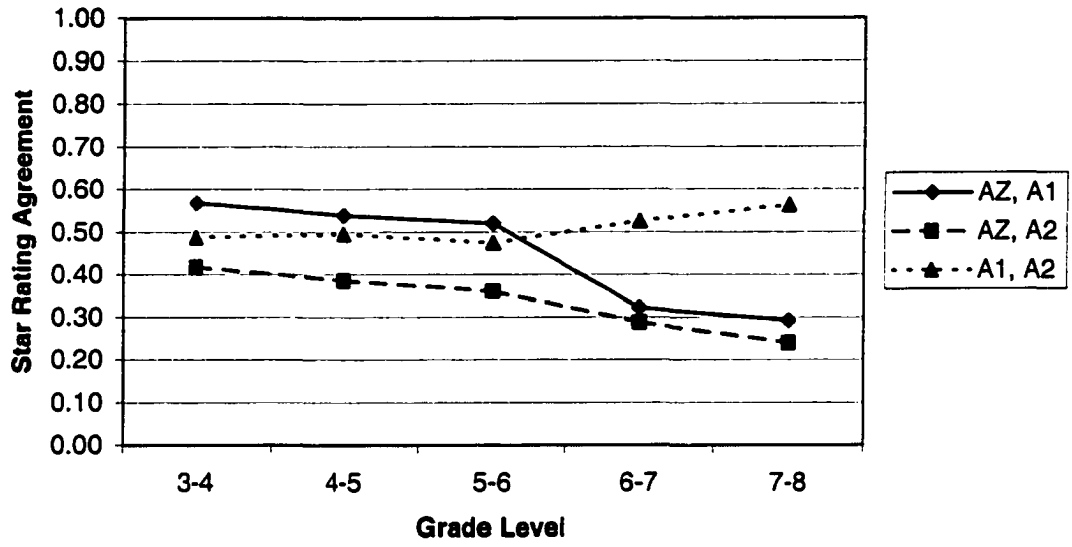

Figure 4.11. Plot of κ of the Star Ratings between each pair of Methods AZ, A1, and A2, by grades in Reading.

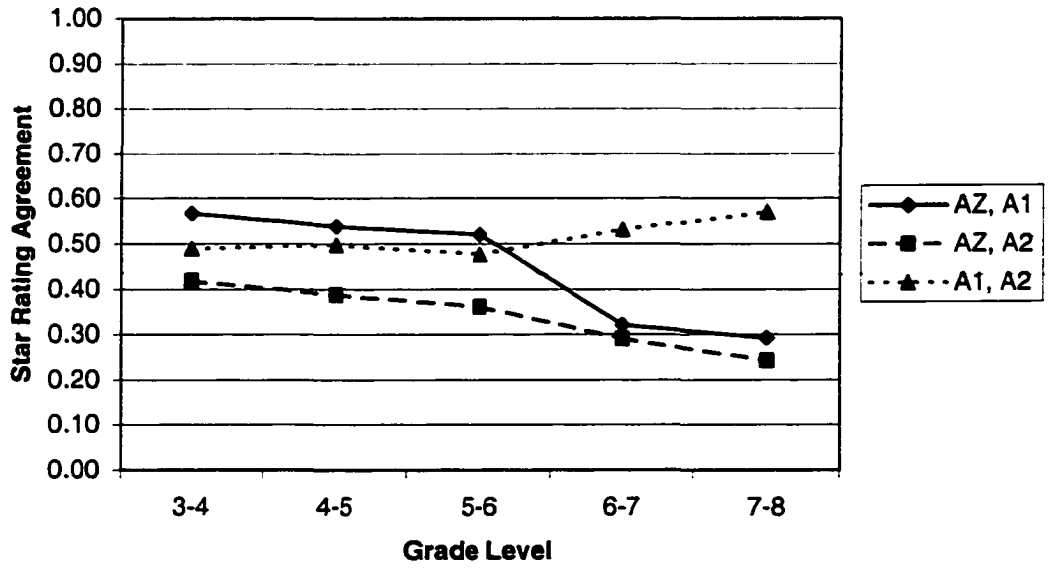

Figure 4.12. Plot of κ / κ_{max} of the Star Ratings between each pair of Methods AZ, A1, and A2, by grades in Reading.

Table 4.24

κ, κ_{max}, and κ / κ_{max} of the Star Ratings between each Pair of Methods AZ, A1, and A2, by Grades, Mathematics

Grade Level	AZ and A1			AZ and A2			A1 and A2		
	κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}
3-4	.419	1.00	.419	.315	.988	.319	.624	.988	.631
4-5	.590	1.00	.590	.476	.998	.447	.601	.998	.602
5-6	.321	1.00	.321	.287	.992	.290	.563	.992	.568
6-7	.580	1.00	.580	.458	1.00	.458	.590	1.00	.590
7-8	.326	1.00	.326	.300	1.00	.300	.541	1.00	.541

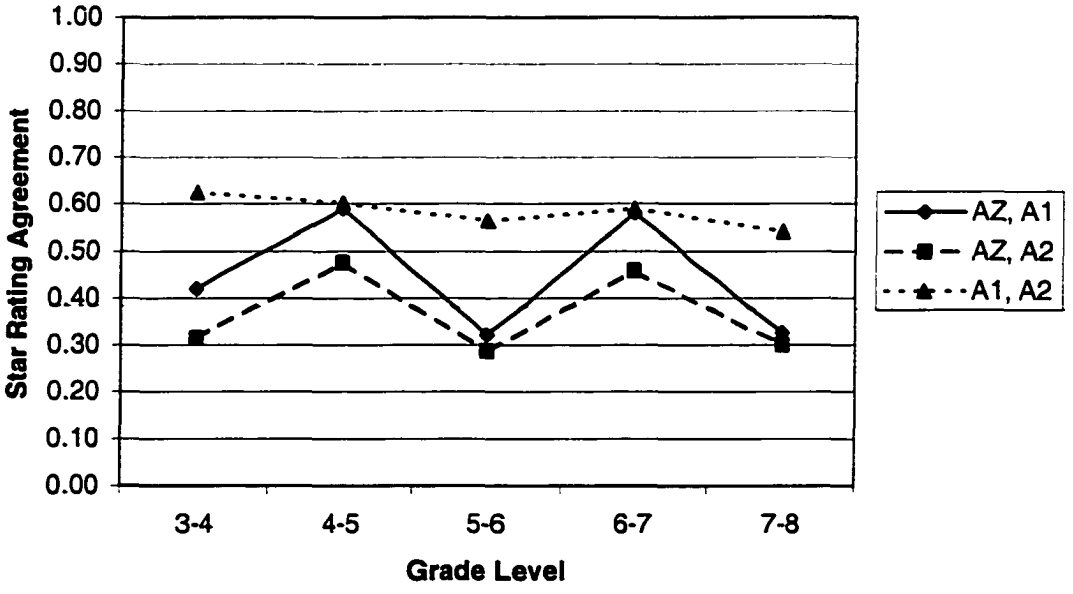

Figure 4.13. Plot of κ of the Star Ratings between each pair of Methods AZ, A1, and A2, by grades in Mathematics.

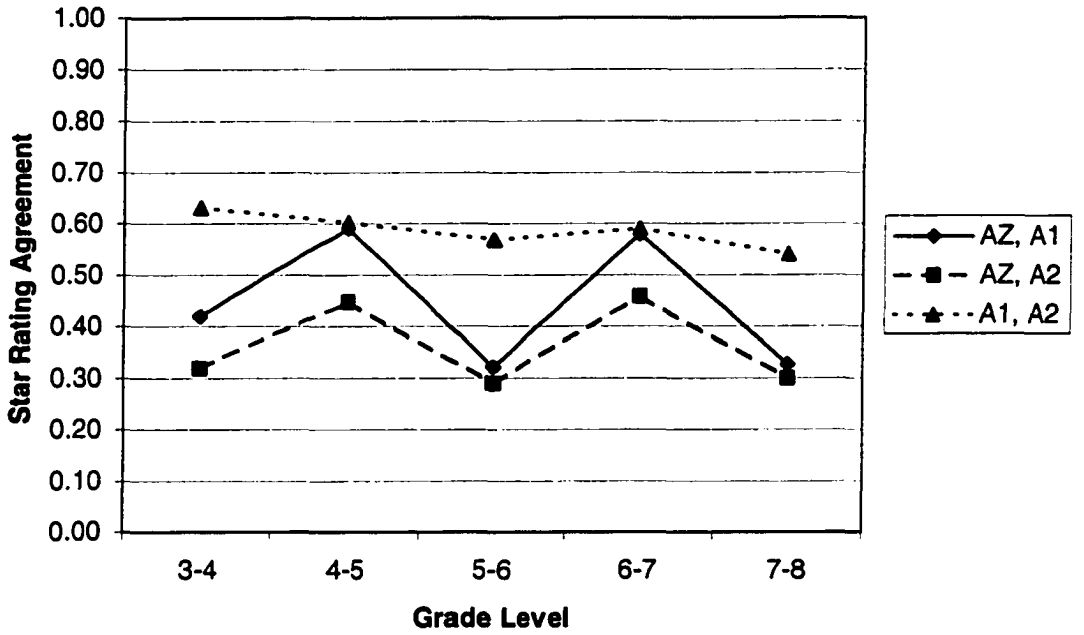

Figure 4.14. Plot of κ / κ_{max} of the Star Ratings between each pair of Methods AZ, A1, and A2, by grades in Mathematics.

Tables 4.23 – 4.24 and Figures 4.11 – 4.14 suggest that the agreement of the Star Ratings between methods is lower overall than for the OYG decisions. The Star Ratings are affected by method choice more than OYG because there are more opportunities for disagreement with the 5 categories of the Star Rating. In addition to the lower agreement between methods, the κ / κ_{max} indicators are lower than for OYG. Despite the overall lower agreement, the agreement between Methods A1 and A2 is still slightly higher than for other pairing.

Tables 4.23 and 4.24 contain a number of κ_{max} values at or near 1.00. To help explain these large values, the contingency table of the Star Rating agreement between Methods AZ and A1 for grade level 3-4 is provided in Table 4.25.

Table 4.25

Star Rating Agreement between Methods AZ and A1 Grade 3-4, Mathematics

Star Rating (Method AZ)		Star Rating (Method A1) 1	2	3	4	5	Total
1	Count	89	31	8	1	0	129
2	Count	37	49	30	14	0	130
3	Count	3	42	53	29	3	130
4	Count	0	7	35	59	29	130
5	Count	0	1	4	27	97	129
Total	Count	129	130	130	130	129	648

The marginal totals in Table 4.25 show that each of the methods assigned the same number of units to each Star Rating category. This is not a coincidence as the Star Ratings are constructed by making five equally numbered groups. With identical marginal values, the κ_{max} values will be 1.00.

Results for Category II:

The Effects of Inappropriately Correcting for RTM

The analyses from Category II were designed to examined the impact of Arizona's correction for RTM on the OYG decision and Star Rating for each school/grade unit. Plots of the proportions of schools achieving OYG from Methods AZ and AZ_{NC} are provided in Figures 4.15 and 4.16.

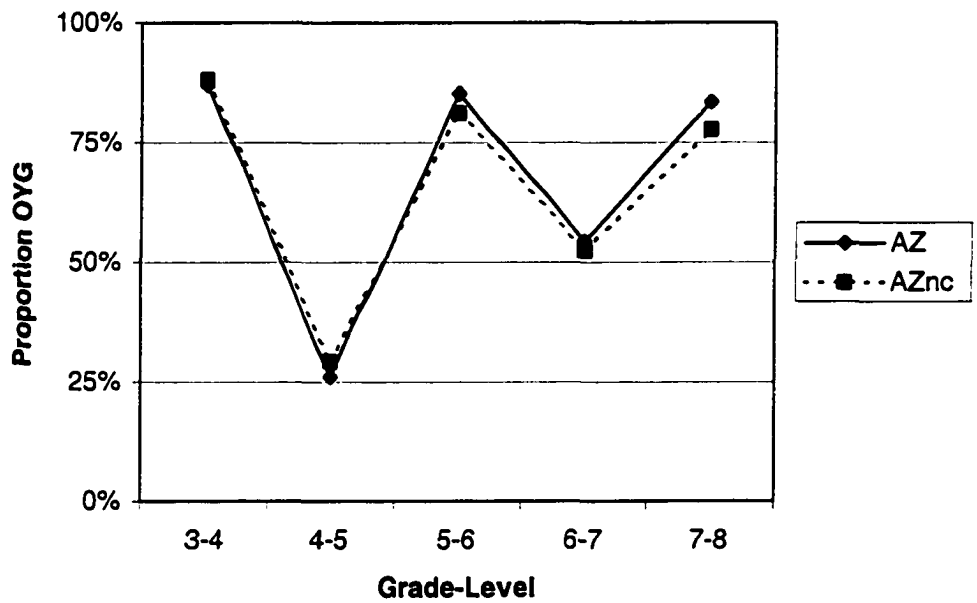

Figure 4.15. Proportions of schools achieving OYG by grade level using Methods AZ and AZ_{NC} in Reading.

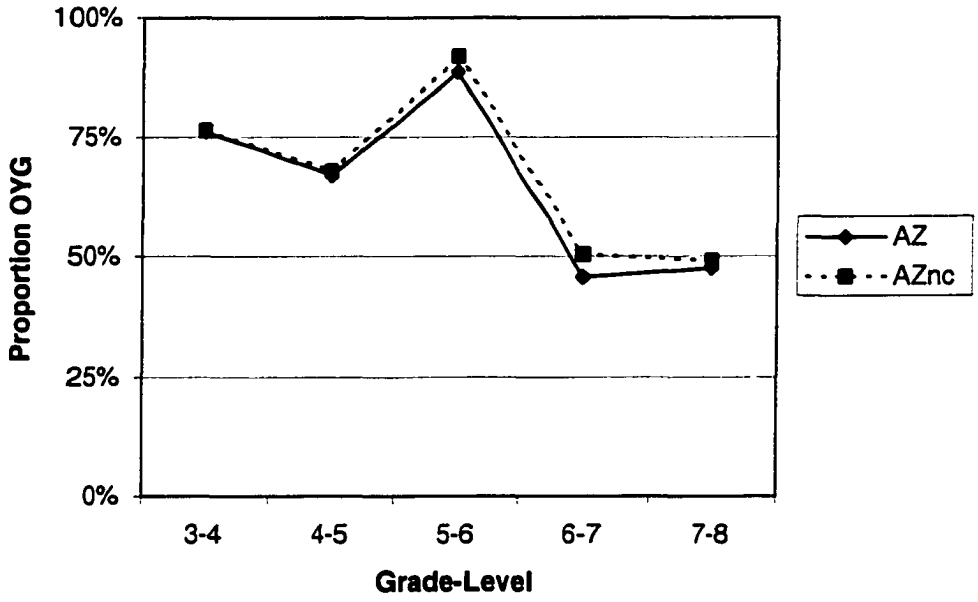

Figure 4.16. Proportions of schools achieving OYG by grade level using Methods AZ and AZ_{NC} in Mathematics.

A review of the plots in Figures 4.15 and 4.16 show few differences between the overall proportion of schools achieving OYG when computed with Method AZ or AZ_{NC}. The proportions appear similar because the correction for RTM causes an increase of schools achieving OYG for those above the mean and an OYG decrease for schools below the mean. The results in the next section demonstrate more clearly the differences between Methods AZ and AZ_{NC}.

Results for Research Question II - 1

These analyses examined the agreement between Methods AZ, and AZ_{NC} in the assignment of OYG to schools grouped by grade level. The results of the κ and κ / κ_{max} analyses on the agreement of the OYG decisions for school/grade units using Methods

AZ and AZ_{NC} are presented in the Table 4.26 and Figure 4.17 for Reading and Table 4.27 and Figure 4.18 for Mathematics.

Table 4.26

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between Methods AZ and AZ_{NC}, by Grades, Reading

Grade Level	AZ and AZ_{NC}		
	κ	κ_{max}	κ / κ_{max}
3-4	.709	.950	.746
4-5	.648	.921	.704
5-6	.600	.856	.701
6-7	.773	.961	.804
7-8	.435	.818	.532

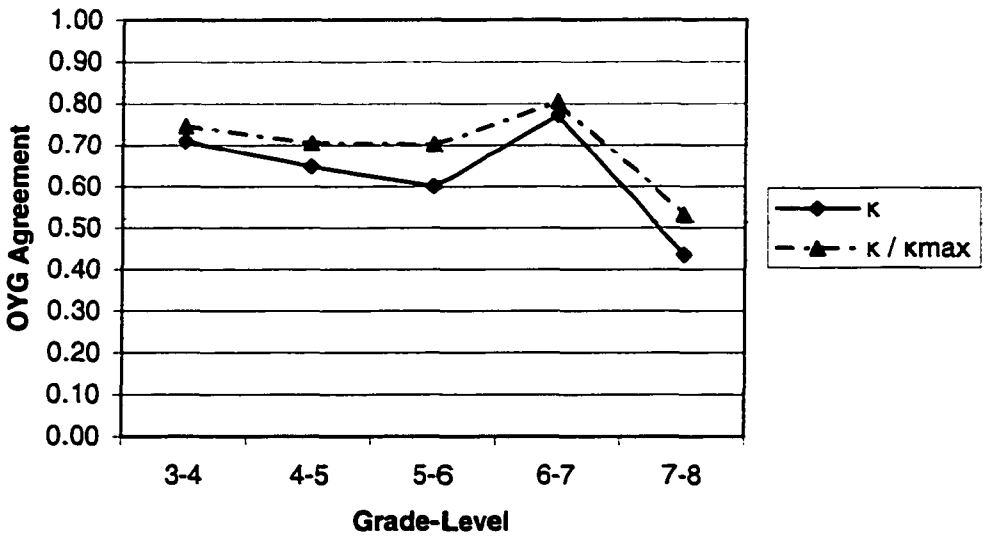

Figure 4.17. Plot of κ and κ / κ_{max} of the OYG decision between Methods AZ and AZ_{NC}, by grades in Reading.

Table 4.27

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between Methods AZ and AZ_{NC}, by Grades, Mathematics

Grade Level	AZ and AZ_{NC}		
	κ	κ_{max}	κ / κ_{max}
3-4	.663	.996	.666
4-5	.684	.979	.698
5-6	.759	.810	.937
6-7	.719	.906	.793
7-8	.680	.970	.702

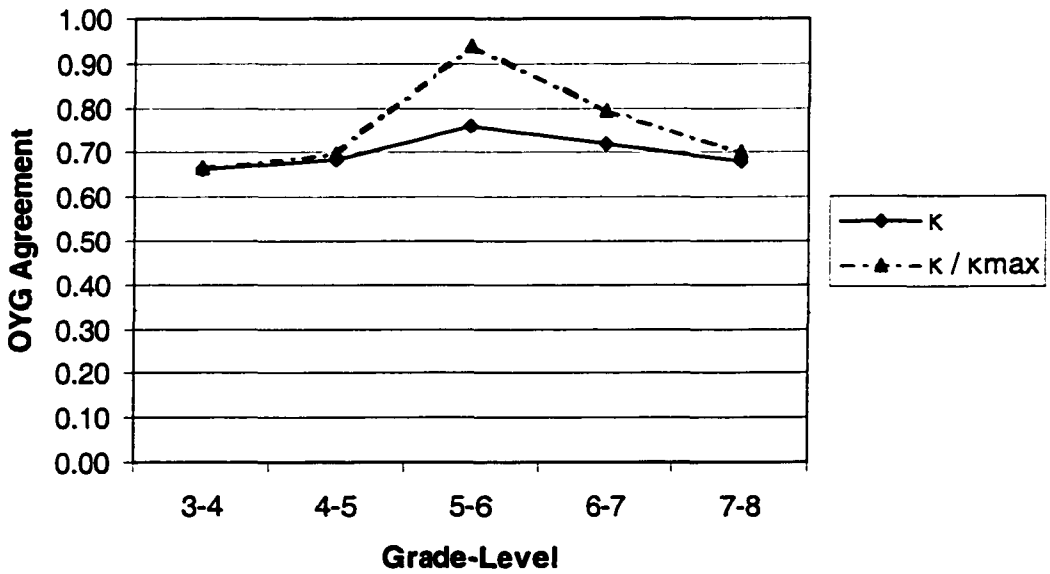

Figure 4.18. Plot of κ and κ / κ_{max} of the OYG decision between Methods AZ and AZ_{NC}, by grades in Mathematics.

The agreement (κ) from Tables 4.26 – 4.27 and Figures 4.17 – 4.18 are not as high as might be expected given the similarities of the overall proportions seen in Figures 4.15 and 4.16. In particular for grades 7-8 in Reading κ is .435 and the κ / κ_{max} is .532. Despite the similarities from the overall proportions, the two methods are not identifying many of the same schools as achieving OYG. The lack of agreement between Methods AZ and AZ_{NC} is examined in more detail later in research question II-3.

Results for Research Question II – 2

These analyses examined the agreement between Methods AZ, and AZ_{NC} in the ranking of growth (the Star Rating) grouped by grade level. The results of the κ and κ / κ_{max} analyses on the agreement of the Star Ratings are presented in the Table 28 and Figure 4.19 for Reading and Table 4.29 and Figure 4.20 for Mathematics.

Table 4.28

κ, κ_{max}, and κ / κ_{max} of the Star Ratings between Methods AZ and AZ_{NC}, by Grades, Reading

Grade Level	AZ and AZ_{NC}		
	κ	κ_{max}	κ / κ_{max}
3-4	.401	.998	.402
4-5	.344	1.00	.344
5-6	.371	1.00	.371
6-7	.429	1.00	.429
7-8	.300	1.00	.300

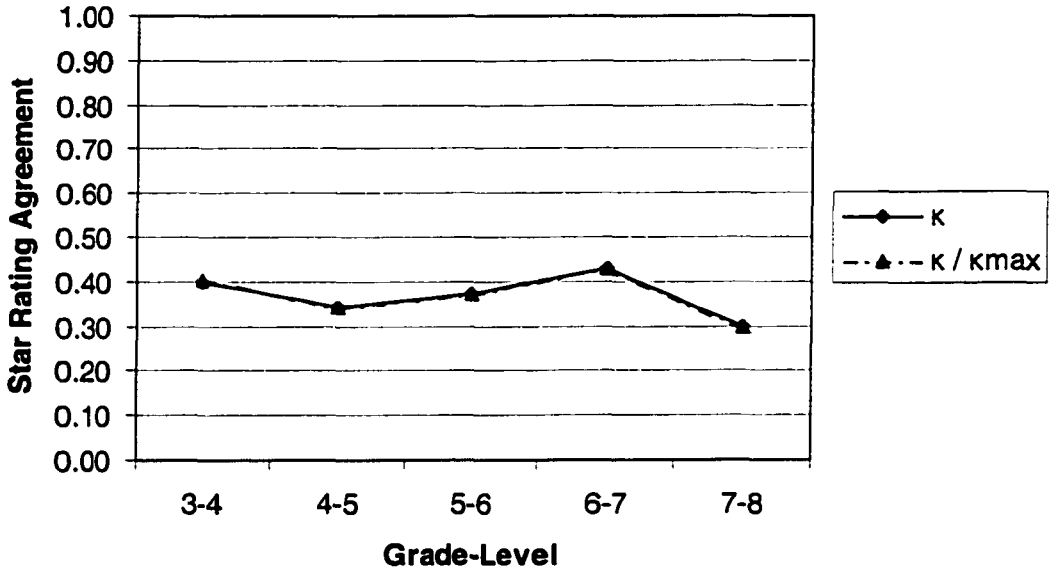

Figure 4.19. Plot of κ and κ / κ_{max} of the Star Ratings between Methods AZ and AZ_{NC}, by grades in Reading.

Table 4.29

κ, κ_{max}, and κ / κ_{max} of the Star Ratings between Methods AZ and AZ_{NC}, by Grades, Mathematics

Grade Level	AZ and AZ_{NC}		
	κ	κ_{max}	κ / κ_{max}
3-4	.394	1.00	.394
4-5	.482	1.00	.482
5-6	.491	1.00	.491
6-7	.463	1.00	.463
7-8	.420	1.00	.420

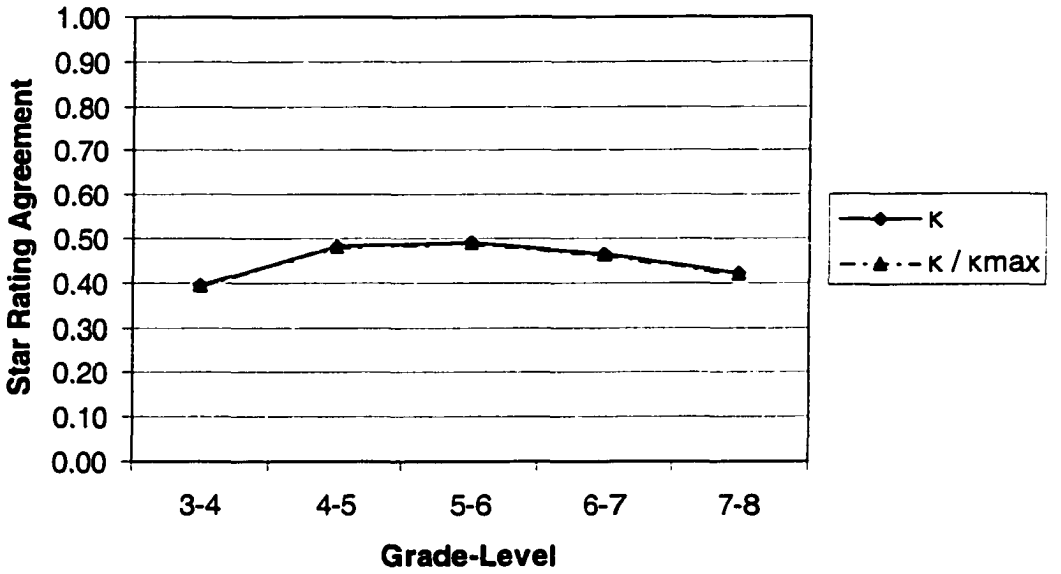

Figure 4.20. Plot of κ and κ / κ_{max} of the Star Ratings between Methods AZ and AZ_{NC}, by grades in Mathematics.

The data in Tables 4.28 – 4.29 and Figures 4.19 – 4.20 show similar results for the Star Ratings as the OYG. Even though the overall proportions of schools achieving OYG is similar between Methods AZ and AZ_{NC} the two methods are not assigning the same Star Rating to many of the schools. While these results suggest differences, the schools need to be regrouped by initial status to demonstrate why these differences occur. The next research question shows these results.

Results for Research Question II - 3

The results of the κ and κ / κ_{max} analyses on the agreement of the OYG decisions by the initial status (the 1998 scaled score mean) groups of the school/grade units using Methods AZ and AZ_{NC} are presented in the Table 4.30 and Figure 4.21 for Reading and Table 4.31 and Figure 4.22 for Mathematics.

Table 4.30

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between Methods AZ and AZ_{NC}, by Grades by Initial Status Groups, Reading

Grade Level	Initial Status Group	Group $\bar{x}$	Grade Level $\bar{x}$ and $\hat{\sigma}$	AZ and AZ_{NC}		
				κ	κ_{max}	κ / κ_{max}
3-4	557-596	582.7	$\bar{x} = 615.6$ $\hat{\sigma} = 21.7$	.434	.434	1.00
	597-611	604.5		.886	.886	1.00
	612-624	617.9		.937	.937	1.00
	625-634	629.5		.730	.730	1.00
	635-662	643.4		.690	.690	1.00
4-5	584-624	611.2	$\bar{x} = 642.8$ $\hat{\sigma} = 21.1$	.483	.483	1.00
	625-638	632.2		.703	.703	1.00
	639-649	644.4		.877	.982	.893
	650-661	655.8		.603	.603	1.00
	662-696	670.2		.437	.437	1.00
5-6	608-639	628.6	$\bar{x} = 656.2$ $\hat{\sigma} = 18.1$	.403	.403	1.00
	640-652	646.9		.815	1.00	.924
	653-663	658.1		.924	.924	1.00
	664-673	668.3		.593	.593	1.00
	674-695	679.1		.327	.327	1.00
6-7	617-646	638.3	$\bar{x} = 663.5$ $\hat{\sigma} = 17.8$	.530	.530	1.00
	647-659	652.8		.919	.919	1.00
	660-669	664.3		.960	.960	1.00
	670-680	674.4		.785	.785	1.00
	681-709	687.9		.499	.499	1.00
7-8	631-662	653.3	$\bar{x} = 681.7$ $\hat{\sigma} = 18.9$	.108	.108	1.00
	663-679	671.6		.422	.422	1.00
	680-688	684.2		.951	.951	1.00
	689-697	693.0		.584	.584	1.00
	698-732	706.5		.220	.220	1.00

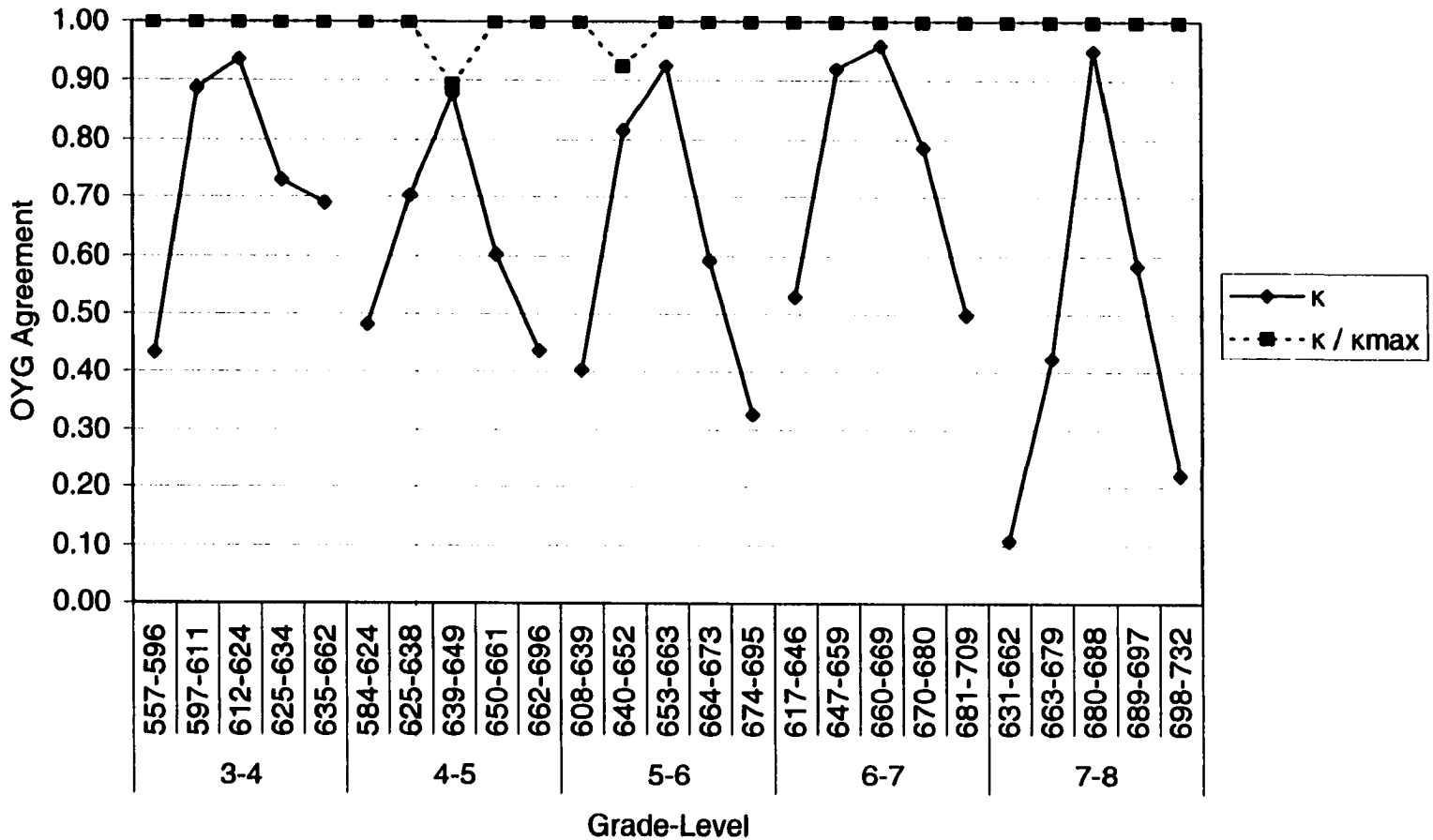

Figure 4.21. Plot of κ and κ / κ_{max} values of the OYG decision between each pair of Methods AZ, and AZ_{NC} by grade by initial status groups in Reading.

Table 4.31

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between Methods AZ and AZ_{NC} by Grades by Initial Status Groups, Mathematics

Grade Level	Initial Status Group	Group $\bar{x}$	Grade Level $\bar{x}$ and $\hat{\sigma}$	AZ and AZ_{NC}		
				κ	κ_{max}	κ / κ_{max}
3-4	557-578	656.9	$\bar{x} = 597.7$ $\hat{\sigma} = 21.6$	.325	.325	1.00
	579-593	586.5		.809	.809	1.00
	594-609	599.1		.934	.978	.955
	605-615	610.1		.876	.876	1.00
	616-657	626.7		.436	.436	1.00
4-5	568-609	598.7	$\bar{x} = 627.5$ $\hat{\sigma} = 19.9$	.389	.389	1.00
	610-622	616.4		.811	.811	1.00
	623-633	628.5		.946	.946	1.00
	634-645	639.5		.831	.831	1.00
	646-677	654.5		.499	.499	1.00
5-6	597-632	621.8	$\bar{x} = 649.4$ $\hat{\sigma} = 19.5$	.204	.204	1.00
	633-643	638.4		.809	.809	1.00
	644-654	649.4		1.00	1.00	1.00
	655-667	660.4		1.00	1.00	1.00
	668-695	676.8		.789	.789	1.00
6-7	616-641	633.1	$\bar{x} = 661.7$ $\hat{\sigma} = 20.5$	.427	.427	1.00
	642-655	649.2		.727	.727	1.00
	656-668	661.6		.923	1.00	.923
	669-681	674.1		.883	.883	1.00
	682-721	690.5		.691	.691	1.00
7-8	635-658	649.4	$\bar{x} = 675.8$ $\hat{\sigma} = 18.5$	.485	.485	1.00
	659-671	665.4		.736	.736	1.00
	672-680	676.9		.907	1.00	.936
	681-690	685.4		.649	.649	1.00
	691-728	701.9		.674	.674	1.00

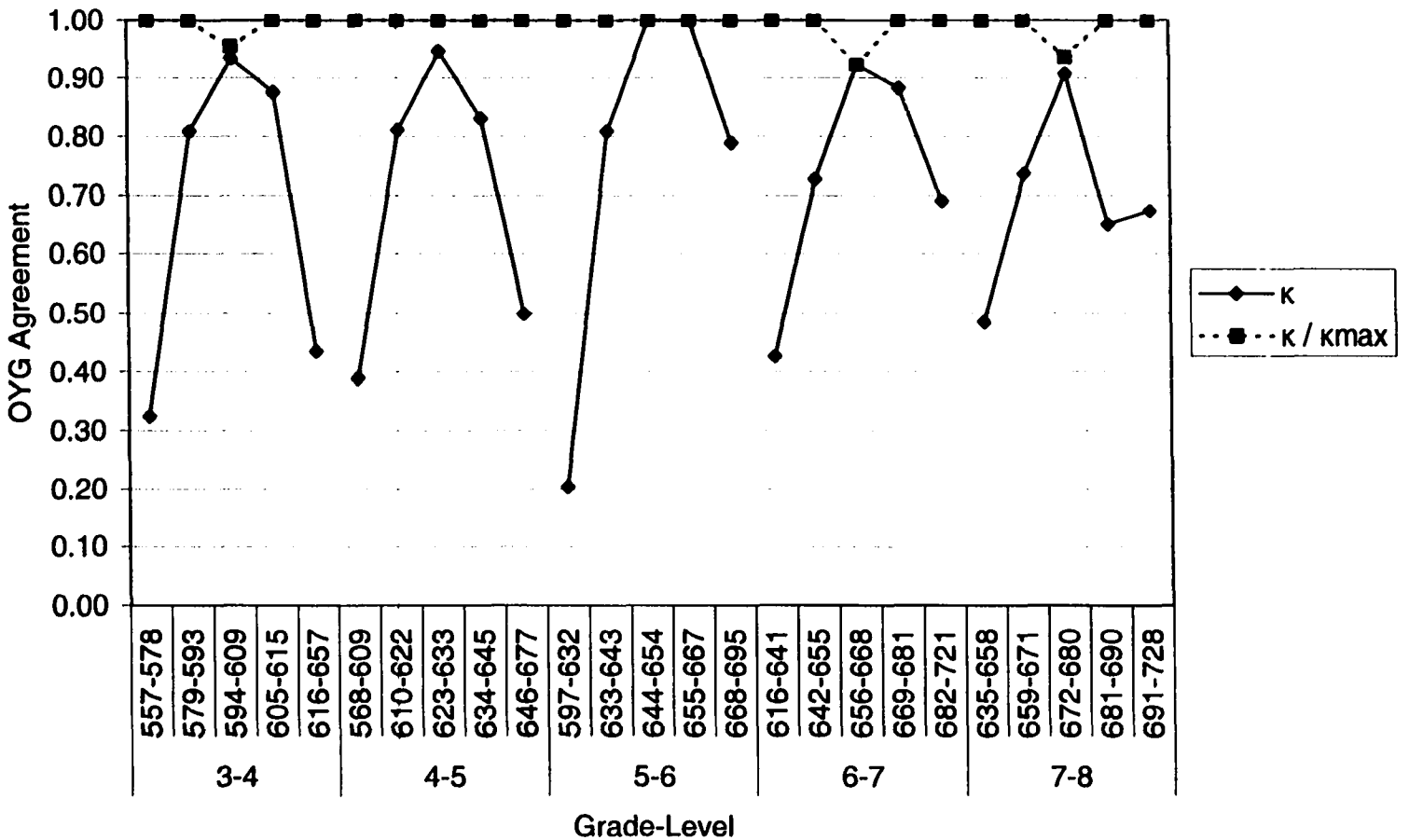

Figure 4.22. Plot of κ and κ / κ_{max} values of the OYG decision between each pair of Methods AZ, and AZ_{NC} by grade by initial status groups in Mathematics.

There is a pattern to the κ values in the plots from Figures 4.21 and 4.22. The agreement between Methods AZ and AZ_{NC} is lower for the initial status groups that are further away from the mean group. The agreement peaks for the middle groups (which include the mean) and then the agreement drops as the initial status of the group moves further above the mean. The pattern is consistent across grade levels and for Reading and Mathematics.

This analysis shows disagreement between Methods AZ and AZ_{NC}. To follow up on the direction of the disagreement, a contingency table showing the counts of school/grade units achieving OYG between the two methods is provided. Table 4.32 shows counts of the OYG decisions for Reading at the 4-5 grade level. This particular area by grade level was selected because it is representative of the other grade levels in the analysis. The table is broken down by initial status groups. In addition, a z-score has been computed for each group to serve as an indicator of distance the mean of each group is away from the overall mean for that particular grade level.

The counts in Table 4.32 show that of the schools in the initial status group with a mean z-score of –1.50, there were 36 that did not achieve OYG with Method AZ but would have if the correction for RTM had not been used (as in Method AZ_{NC}). In the next group, which is closer to the overall mean at -.50, there were 17 schools that did not achieve OYG under Method AZ. As the initial status groups moved above the mean for the grade level, the last two groups, there were 18 and 15 schools respectively, that achieved OYG with Method AZ, but would not have if the RTM correction had not been used. Because of the correction used in Method AZ, it was easier for a unit with an initial

status above the mean to achieve OYG and more difficult for a unit with initial status below the mean.

Table 4.32

Counts of the OYG Decisions between Methods AZ and AZ_{NC} by Initial Status Groups, Reading Grade Level 4-5

Initial Status Group	z-score for Group Mean			AZ_{NC} 0	AZ_{NC} 1
584-624	-1.50	AZ	0	55	36
			1	0	40
625-638	-.50	AZ	0	83	17
			1	0	32
639-649	.08	AZ	0	87	4
			1	3	38
650-661	.62	AZ	0	95	0
			1	18	19
662-696	1.29	AZ	0	110	0
			1	15	7

From Tables 4.30 and 4.31 there are a number of κ / κ_{max} values of 1.00 meaning that the agreement between AZ and AZ_{NC} was perfect given the maximum possible. The counts in Table 4.32 can help to explain these large values. For example, the agreement counts for the first initial status group (with a group z-score of –1.50), shows that Method AZ_{NC} assigns OYG to all the same units that Method AZ does, plus additional units. This decreases the maximum possible agreement (κ) and increases the κ / κ_{max} ratio.

Results for Category III:

The Independence of the Change Indicators

The analyses from Category III examined the relationship between the initial status of the school/grade units and the growth indicators computed in Methods AZ, AZ_{NC}, A1, A2, and A3. Method AZ_{NC} is included in this analysis only for the purposes comparison with Method AZ and not intented as a plausible alternative. To examine for a linear relationship between initial status and amount of growth, scatter-plots are provided from Method AZ for grade level 3-4 (Figure 4.23) and grade level 6-7 (Figure 4.24). Rather than show the plots from all grade levels and methods, these two grade levels were selected because the shapes are fairly representative of the remaining.

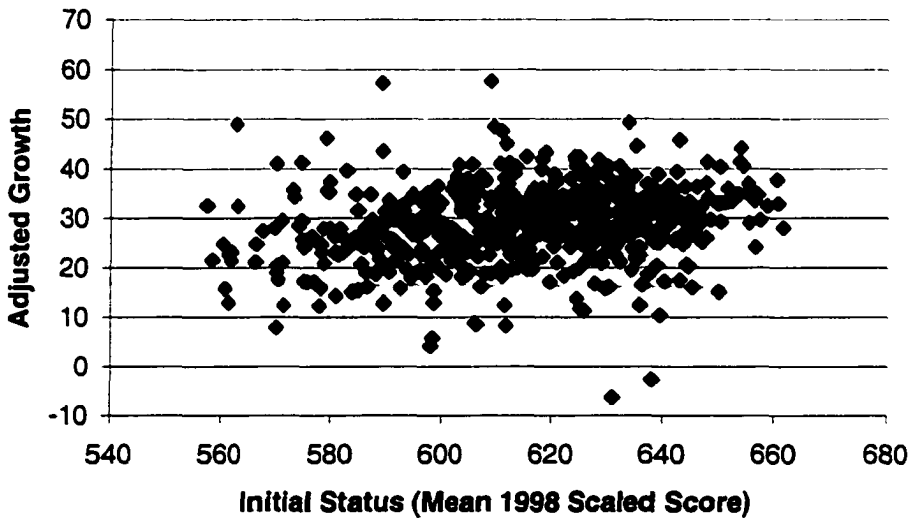

Figure 4.23. Scatter-plot of adjusted growth and initial status for school/grade units using Methods AZ in Reading at grade level 3-4.

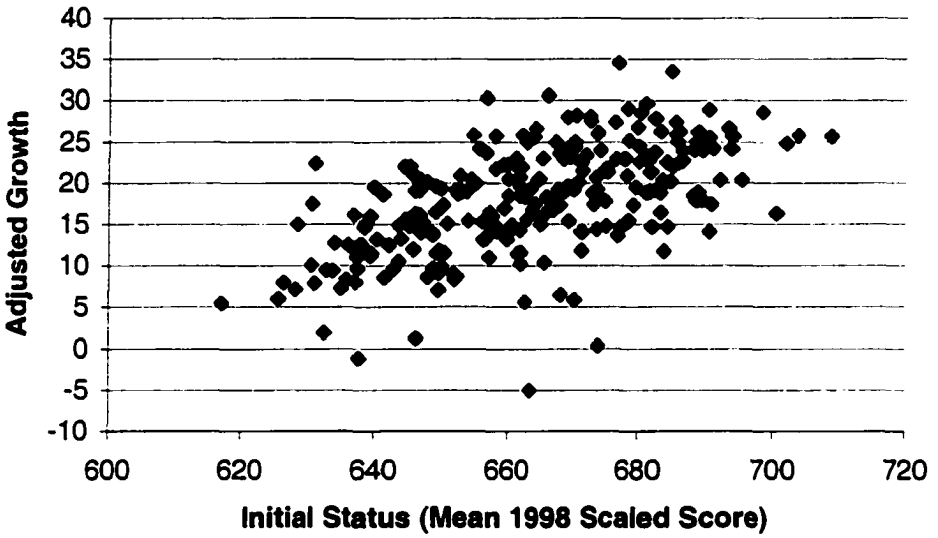

Figure 4.24. Scatter-plot of adjusted growth and initial status for school/grade units using Method AZ in Reading at grade level 6-7.

The shapes of the plots in Figures 4.23 and 4.24 do not suggest a curvilinear relationship between initial status and the amount of growth made by school/grade units. A linear model (such as Pearson's *r*) would seem to be an appropriate analysis.

Results for Research Question III - 1

The Pearson correlations between the growth indicators and initial status from each method are listed in Table 4.33 for Reading and Table 4.34 for Mathematics. The correlations are plotted by grade level in Figures 4.25 and 4.26.

Table 4.33

Correlations between Amount of Growth Indicators and Initial Status for School/Grade Units Using Methods AZ, AZ_{NC}, A1, A2, and A3: Reading

	Method				
Grade Level	AZ	AZ_{NC}	A1	A2	A3
3-4	.234	-.242	-.050	-.014	-.040
4-5	-.019	-.523	-.324	-.286	-.356
5-6	-.065	-.496	-.378	-.312	-.292
6-7	.566	.192	.035	.059	-.121
7-8	-.003	-.537	-.539	-.509	-.549

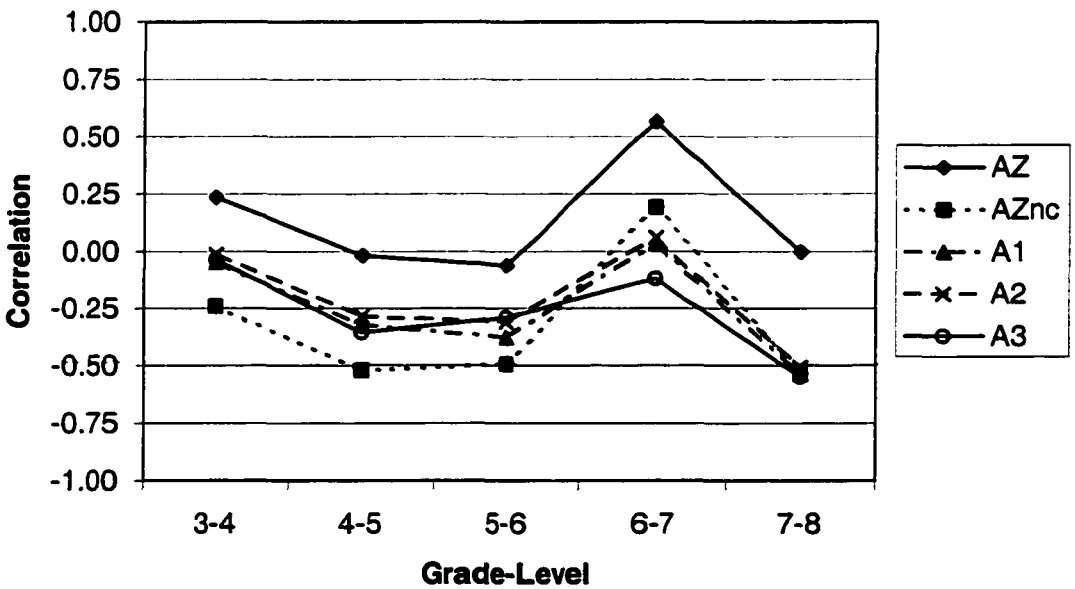

Figure 4.25. Plots of the correlations between amount of growth indicators and initial status for school/grade units using Methods AZ, AZ_{NC}, A1, A2, and A3 in Reading.

Table 4.34

Correlations between Amount of Growth Indicators and Initial Status for School/Grade Units Using Methods AZ, AZ_{NC}, A1, A2, and A3: Mathematics

	Method				
Grade Level	AZ	AZ_{NC}	A1	A2	A3
3-4	.051	-.373	-.349	-.344	-.299
4-5	.139	-.281	-.167	-.141	-.158
5-6	.267	-.104	-.291	-.248	-.266
6-7	.121	-.240	-.132	-.113	-.132
7-8	.191	-.187	-.298	-.224	-.200

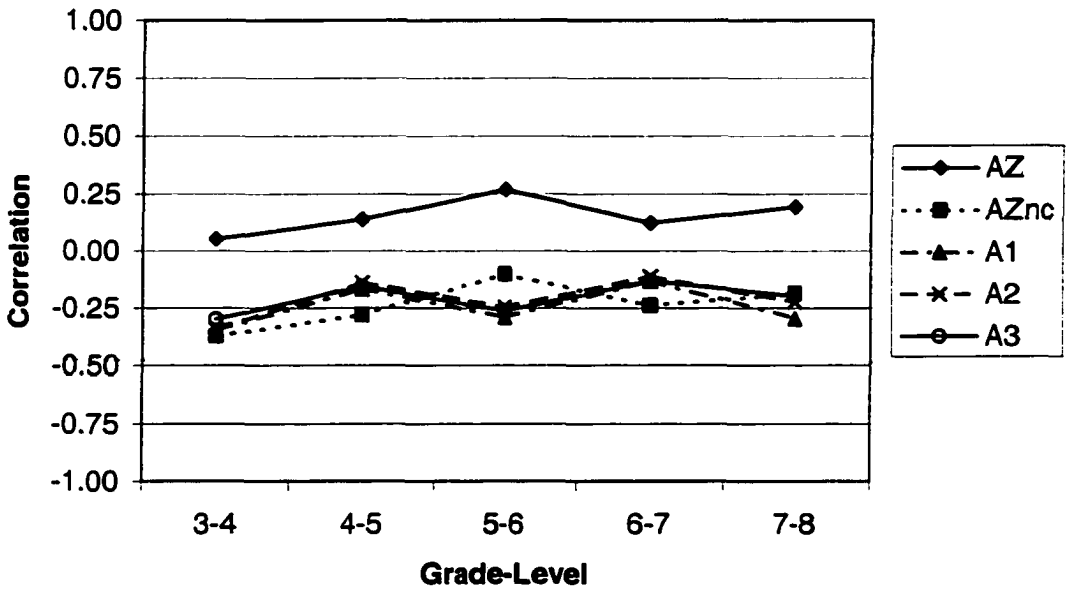

Figure 4.26. Plots of the correlations between amount of growth indicators and initial status for school/grade units using Methods AZ, AZ_{NC}, A1, A2, and A3 in Mathematics.

The impact of correcting for RTM can be seen by comparing the results of Methods AZ and AZ_{NC} in Tables 4.33 and 4.34 and Figures 4.25 and 4.26. There is a systematic decrease in the values of *r* from Method AZ to Method AZ_{NC} across grade levels. In some grade levels, such as 4-5, 6-7, and 7-8, the correction for RTM reduced the magnitude of the correlations to nearly zero. In grade level 6-7 for Reading, the correction increased the correlation. The effect is not systematic across grade levels or subject areas and not adequately explained by the RTM correction in Method AZ.

The plots of the correlations for Reading in Figure 4.25 show that the independence of the amount of growth indicators from Methods A1, A2, and A3 vary across grade levels. While there is grade level variation, all three of these methods are consistent with respect to each other, suggesting grade level effects rather than method dependency. The plots in Figure 4.26 (for Mathematics) suggest the same consistency between Methods A1, A2, and A3 as seen in reading.

In general, the correlations between the amount of growth and initial status are relatively small in magnitude and negative in direction for Methods A1, A2, and A3. The only exception to this is for grade level 7-8 in Reading where the correlations are near -.50. All three of the methods (A1, A2, and A3) show this moderate correlation, again, suggesting a grade level by subject area effect rather than a specific method effect.

Results for Research Question III - 2

The results of the correlations (r_{pb}) between the school OYG indicator and initial status from each method are listed in Table 4.35 for Reading and Table 4.36 for Mathematics. The correlations are plotted by grade level in Figures 4.27 and 4.28.

Table 4.35

Correlations between the OYG Indicators and Initial Status for School/Grade Units Using Methods AZ, AZ_{NC}, A1, A2, and A3: Reading

	Method				
Grade Level	AZ	AZ_{NC}	A1	A2	A3
3-4	.180	-.064	.066	.054	.004
4-5	-.084	-.424	-.277	-.252	-.024
5-6	.075	-.275	-.174	-.125	-.065
6-7	.468	.215	.105	.078	-.049
7-8	.055	-.435	-.468	-.382	-.153

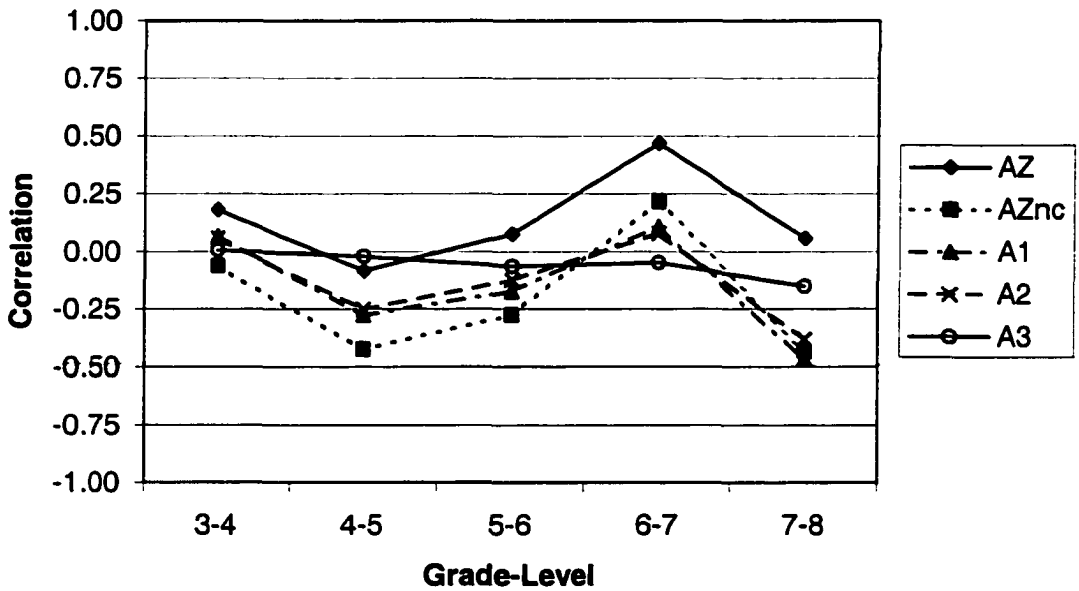

Figure 4.27. Plots of the correlations between the OYG indicators and initial status for school/grade units using Methods AZ, AZ_{NC}, A1, A2, and A3 in Reading.

Table 4.36

Correlations between the OYG Indicators and Initial Status for School/Grade Units Using Methods AZ, AZ_{NC}, A1, A2, and A3: Mathematics

	Method				
Grade Level	AZ	AZ_{NC}	A1	A2	A3
3-4	.101	-.264	-.256	-.218	-.018
4-5	.174	-.195	-.100	-.067	-.070
5-6	.139	-.035	-.155	-.078	-.008
6-7	.109	-.216	-.089	-.080	-.077
7-8	.189	-.133	-.305	-.226	-.095

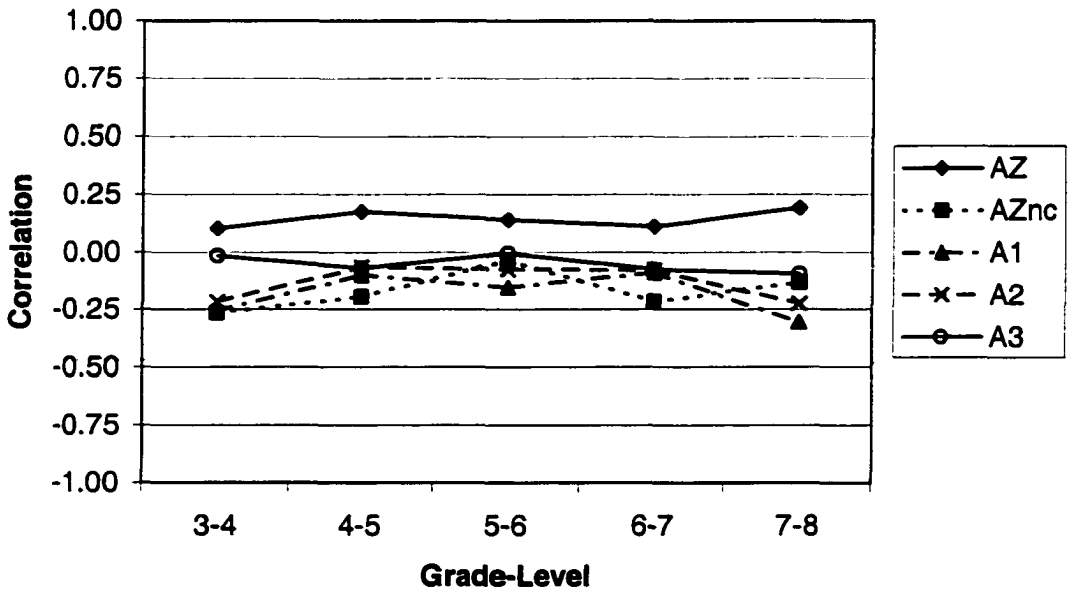

Figure 4.28. Plots of the correlations between the OYG indicators and initial status for school/grade units using Methods AZ, AZ_{NC}, A1, A2, and A3 in Mathematics.

The plots in Figure 4.27 show the same systematic decrease in the values of r_{pb} from Method AZ to Method AZ_{NC} across grade levels in Reading. The magnitudes of the decreases are not as consistent in Mathematics (Figure 4.28), again suggesting the correction for RTM does not have a consistent impact on independence across grade levels or subject areas.

The plots of the correlations for Reading in Figure 4.27 show that the independence of the OYG indicators from Methods A1, A2, and A3 vary across grade levels. Methods A1 and A2 are consistent with respect to each other. The independence of Method A3 shows less variability across grade levels than Methods A1 or A2. A similar relationship between Methods A1, A2, and A3 is evident in Mathematics (Figure 4.28).

Across the plots for research questions III-1 and III-2, Method AZ tends to show a near zero or positive relationship between the growth indicators and initial status, while the other methods tend to show negative relationships. Methods A1, A2, and A3 show similar growth indicator correlations with respect to each other across the analyses. Finally, the OYG growth indicators tend to show less dependence between initial status and amount of growth indicators. This last finding was expected because the OYG indicators have more restricted ranges than the amount of growth indicators. The restricted range has a tendency to reduce the magnitudes of correlations.

Results for Category IV:

The Impact of Accounting for Error in the Three Alternative Methods.

The analyses from Category IV examined the OYG differences between Methods A1 and $A1_{AE}$, A2 and $A2_{AE}$ and $A1_{AE}$, $A2_{AE}$, A3. The results for each of the three research questions in this category are listed separately.

Results for Research Question IV - 1

These analyses examined the agreement between Methods A1 and $A1_{AE}$ in the assignment of OYG to school/grade units grouped by grade level. Plots of the proportions of schools achieving OYG from Methods A1 and $A1_{AE}$ are provided in Figures 4.29 and 4.30.

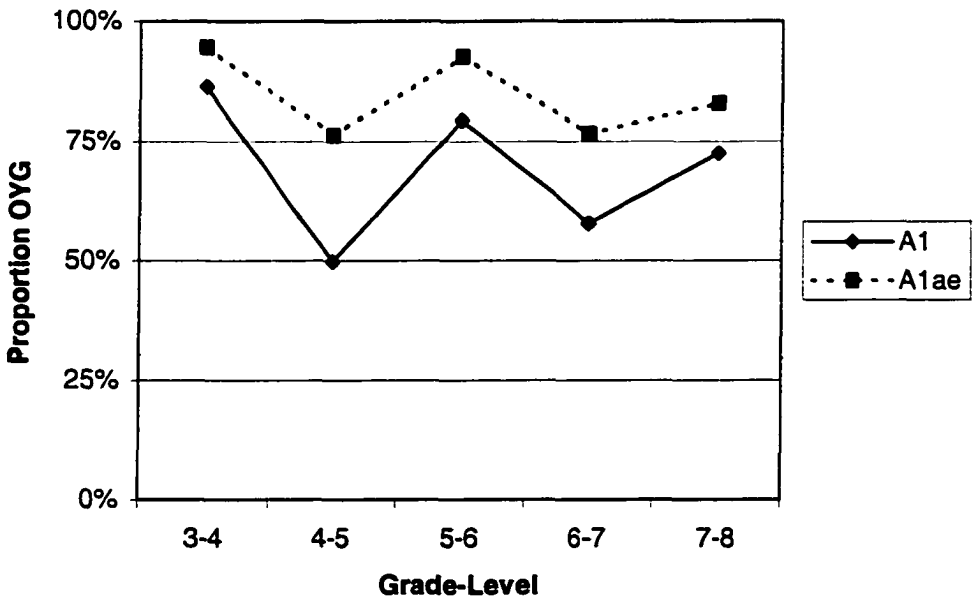

Figure 4.29. Proportions of schools achieving OYG by grade level using Methods A1 and $A1_{AE}$ in Reading.

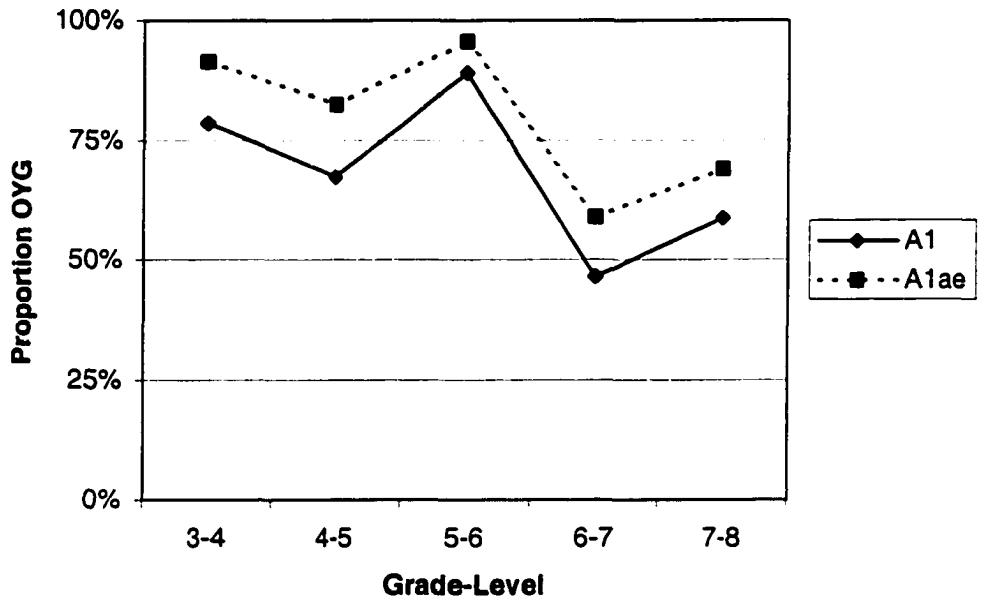

Figure 4.30. Proportions of schools achieving OYG by grade level using Methods A1 and $A1_{AE}$ in Mathematics.

The plots in Figures 4.29 and 4.30 show that the adjustments to account for error in Method A1 increase the proportions of schools achieving OYG at each grade level. The inconsistent proportion increase between Method A1 and $A1_{AE}$ at each grade level is due to the grade level standard errors used in the adjustments.

The results of the κ and κ / κ_{max} analyses on the agreement of the OYG decisions for school/grade units using Methods A1 and $A1_{AE}$ are presented in the Table 4.37 and Figure 4.31 for Reading and Table 4.38 and Figure 4.32 for Mathematics.

Table 4.37

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between Methods A1 and $A1_{AE}$, by Grades, Reading

Grade Level	A1 and $A1_{AE}$		
	κ	κ_{max}	κ / κ_{max}
3-4	.533	.533	1.0
4-5	.470	.470	1.0
5-6	.470	.470	1.0
6-7	.591	.591	1.0
7-8	.709	.709	1.0

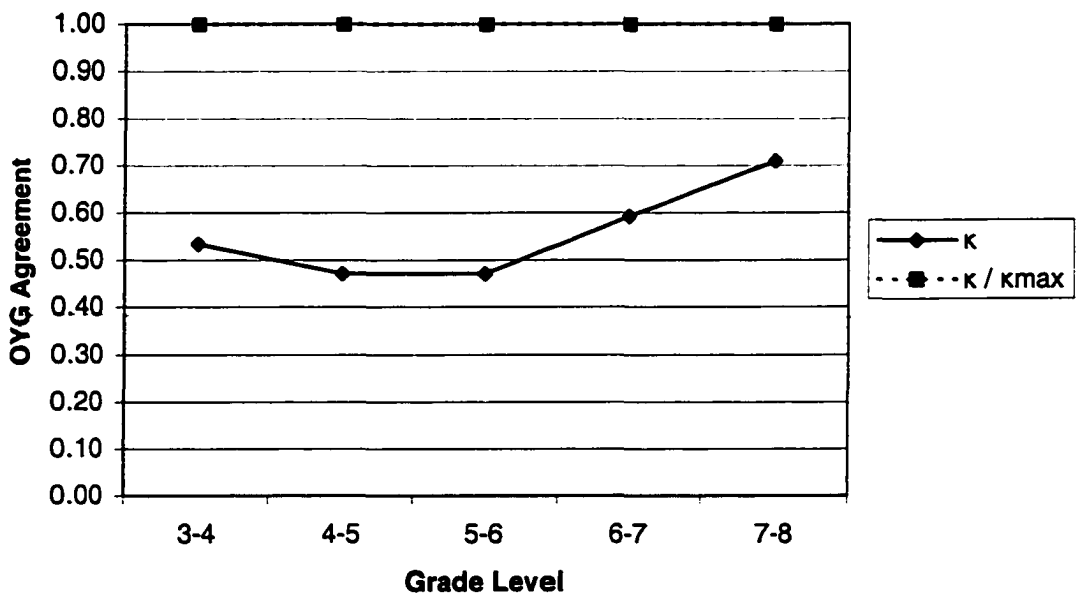

Figure 4.31. Plots of κ and κ / κ_{max} of the OYG decision between Methods A1 and $A1_{AE}$, by grades in Reading.

Table 4.38

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between Methods A1 and $A1_{AE}$, by Grades, Mathematics

Grade Level	A1 and $A1_{AE}$		
	κ	κ_{max}	κ / κ_{max}
3-4	.518	.518	1.0
4-5	.605	.605	1.0
5-6	.551	.551	1.0
6-7	.753	.753	1.0
7-8	.781	.781	1.0

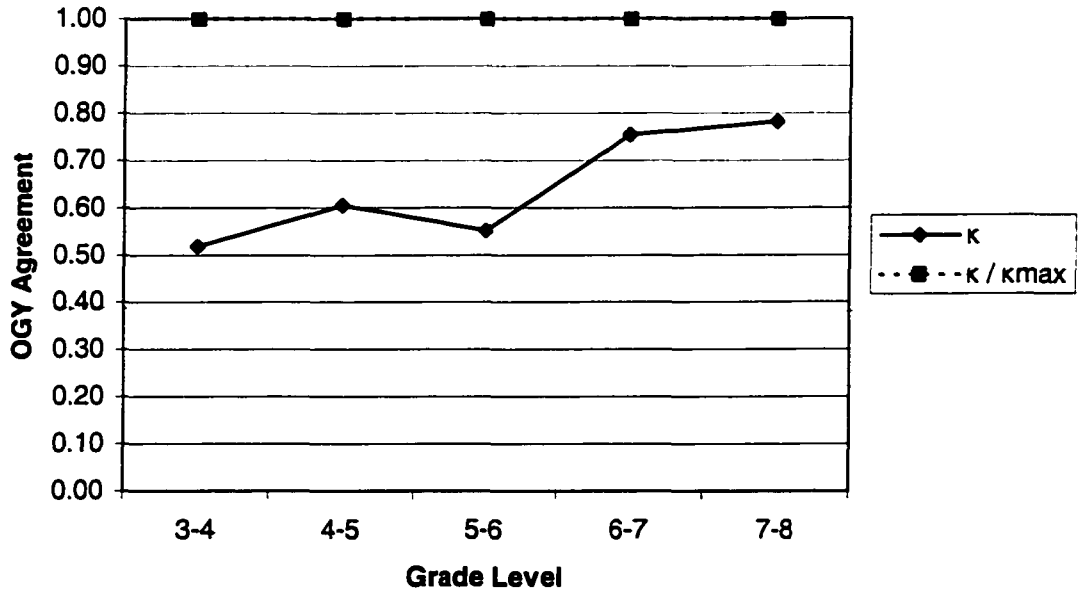

Figure 4.32. Plots of κ and κ / κ_{max} of the OYG decision between Methods A1 and $A1_{AE}$, by grades in Mathematics.

The results from Tables 4.37 and 4.38 and Figures 4.31 and 4.32 show the OYG agreement between A1 and $A1_{AE}$. The two methods did not result in the same number of school/grade units that would achieve OYG, thus the disagreement seen in these results. For both Reading and Mathematics there was a slight increase in the κ values at the later grade levels, suggesting that accounting for error will have a larger impact in the earlier grade levels (3-4, 4-5, and 5-6). The κ / κ_{max} indicators for all the grade levels are 1.00 meaning that the agreement was perfect given the maximum possible. To help explain these large values, the contingency table of the OYG agreement between Methods A1 and $A1_{AE}$ for grade level 7-8 is provided in Table 4.39.

Table 4.39

OYG Agreement between Methods A1 and $A1_{AE}$ Grade 7-8, Reading

		Method $A1_{AE}$		
Method A1		0	1	Total
0	Count	35	53	88
1	Count	0	555	555
Total	Count	35	608	643

The counts in Table 4.39 shows that every school identified as achieving OYG by Method A1 was also identified as achieving OYG by Method $A1_{AE}$. These results make sense, as the adjustments in Method $A1_{AE}$ result in a constant (within a grade level) being subtracted from the OYG criteria. The constant results in Method $A1_{AE}$ assigning OYG to all the same units that Method A1 does, plus a few additional units. This decreases the maximum possible agreement and increases the κ / κ_{max} ratio.

Results for Research Question IV - 2

This question examined the agreement between Methods A2 and $A2_{AE}$ in the assignment of OYG to school/grade units grouped by grade level. Plots of the proportions of schools achieving OYG from Methods A1 and $A1_{AE}$ are provided in Figures 4.33 and 4.34.

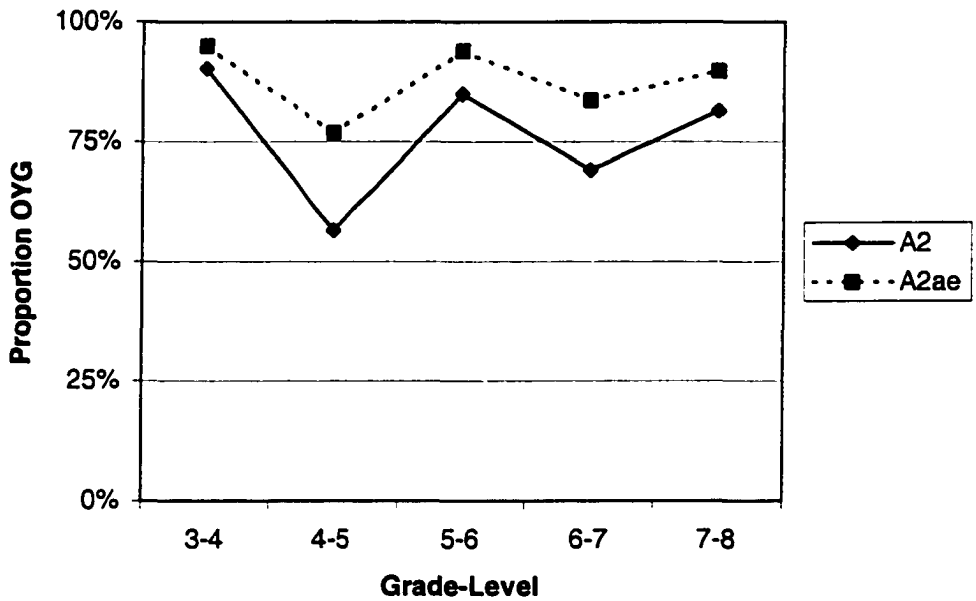

Figure 4.33. Proportions of schools achieving OYG by grade level using Methods A2 and $A2_{AE}$ in Reading.

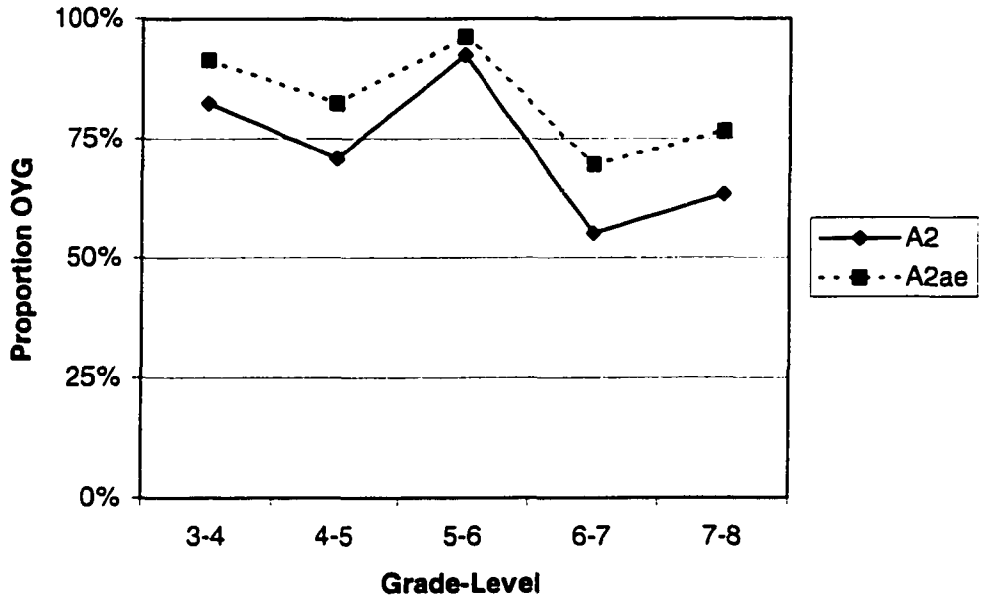

Figure 4.34. Proportions of schools achieving OYG by grade level using Methods A2 and $A2_{AE}$ in Mathematics.

The plots in Figures 4.33 and 4.34 show that the adjustments to account for error in Method A2 increase the proportions of schools achieving OYG at each grade level. The resulting increase in the proportion of schools achieving OYG from Method $A2_{AE}$ over that of Method A2 was approximately the same as the increase in Method $A1_{AE}$ over A1. As mentioned earlier, the increases are not consistent across grades because the standard errors are different (computed separately) for each grade level.

The results of the κ and κ / κ_{max} analyses on the agreement of the OYG decisions for school/grade units using Methods A2 and $A2_{AE}$ are presented in the Table 4.40 and Figure 4.35 for Reading and Table 4.41 and Figure 4.36 for Mathematics.

Table 4.40

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between Methods A2 and $A2_{AE}$, by Grades, Reading

Grade Level	A2 and $A2_{AE}$		
	κ	κ_{max}	κ / κ_{max}
3-4	.665	.665	1.0
4-5	.563	.563	1.0
5-6	.534	.534	1.0
6-7	.611	.611	1.0
7-8	.664	.664	1.0

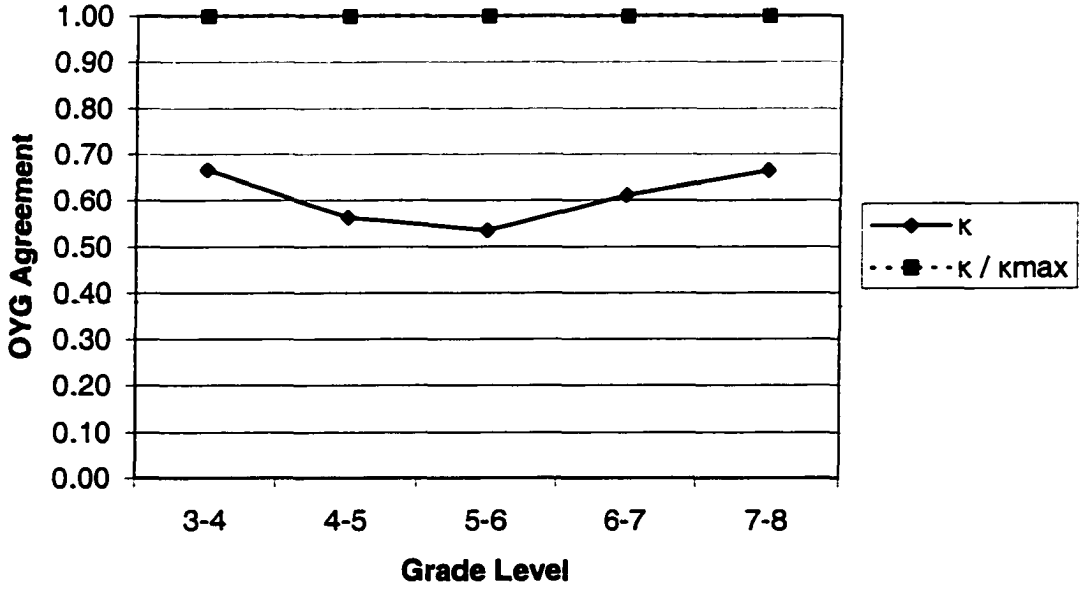

Figure 4.35. Plots of κ, and κ / κ_{max} of the OYG decision between Methods A2 and $A2_{AE}$, by grades in Reading.

Table 4.41

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between Methods A2 and $A2_{AE}$, by Grades, Mathematics

Grade Level	A2 and $A2_{AE}$		
	κ	κ_{max}	κ / κ_{max}
3-4	.614	.614	1.0
4-5	.681	.681	1.0
5-6	.649	.649	1.0
6-7	.699	.699	1.0
7-8	.692	.692	1.0

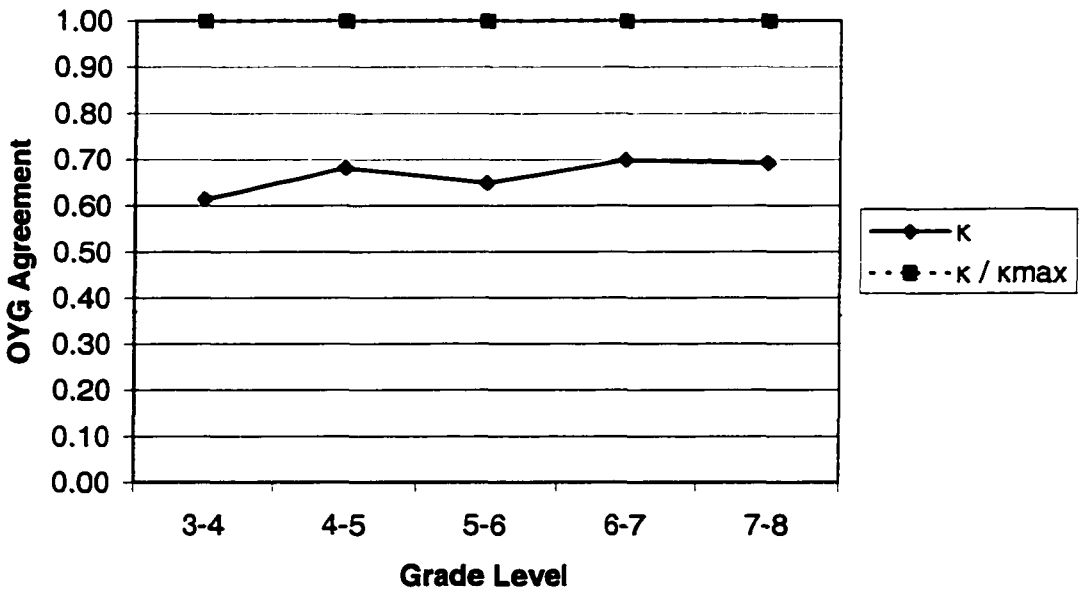

Figure 4.36. Plots of κ, and κ / κ_{max} of the OYG decision between Methods A2 and $A2_{AE}$, by grades in Mathematics.

The values from Table 4.40 and Figure 4.35 suggest that accounting for error in Method A2 has a more of an impact in grade levels 4-5 and 5-6 for Reading. There does not appear to be a specific grade level effect for Mathematics (Table 4.41 and Figure 4.36). Again, the κ / κ_{max} values are 1.00 because the adjustment in $A2_{AE}$ results in a constant reduction of the OYG criterion within a grade level.

Results for Research Question IV - 3

This question examined the agreement between Methods $A1_{AE}$, $A2_{AE}$ and A3 in the assignment of OYG to school/grade units grouped by grade level. Plots of the proportions of schools achieving OYG from these Methods are provided in Figures 4.37 and 38.

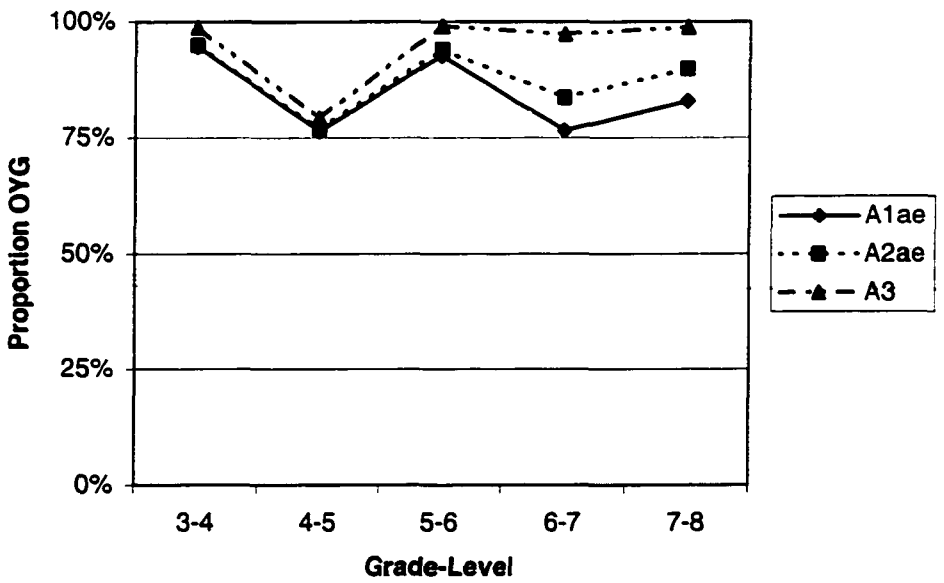

Figure 4.37. Proportions of schools achieving OYG by grade level using Methods $A1_{AE}$, $A2_{AE}$, and A3 in Reading.

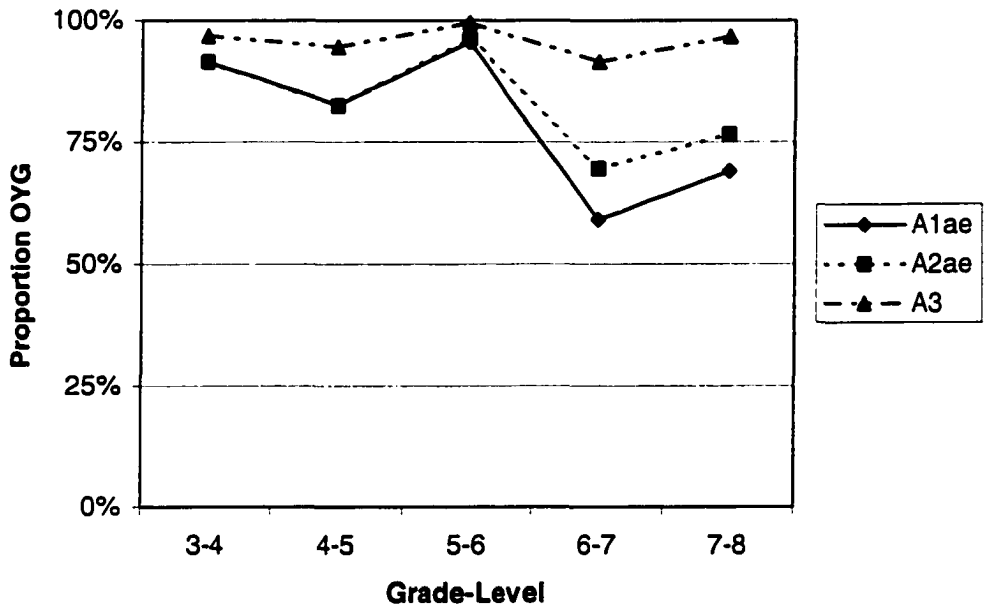

Figure 4.38. Proportions of schools achieving OYG by grade level using Methods $A1_{AE}$, $A2_{AE}$, and A3 in Mathematics.

The plots in Figures 4.37 and 4.38 show that Method A3 produced the highest proportions of schools achieving OYG. Method $A2_{AE}$ produces the second highest results and $A1_{AE}$ the lowest. In Reading, Method A3 shows a drop in the 4-5 grade level whereas this drop is not evident in Mathematics. Across both plots, all three methods result in similar proportions of schools achieving OYG at grades 3-4, 4-5, and 5-6. There is a larger difference between the methods for grade levels 6-7 and 7-8.

The results of the κ and κ / κ_{max} analyses on the agreement of the OYG decisions for school/grade units using Methods $A1_{AE}$, $A2_{AE}$, and A3 are presented in the Table 4.42 and Figures 4.39 - 4.40 for Reading and Table 4.43 and Figures 4.41 – 4.42 for Mathematics.

Table 4.42

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between each Pair of Methods $A1_{AE}$, $A2_{AE}$, and A3, by Grades, Reading

Grade Level	$A1_{AE}$ and $A2_{AE}$			$A1_{AE}$ and A3			$A2_{AE}$ and A3		
	κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}
3-4	.752	.969	.776	.349	.396	.882	.416	.416	1.00
4-5	.683	.987	.692	.175	.175	1.00	.154	.179	.863
5-6	.536	.893	.599	.203	.203	1.00	.246	.246	1.00
6-7	.587	.781	.751	.168	.168	1.00	.251	.251	1.00
7-8	.685	.710	.964	.111	.111	1.00	.193	.193	1.00

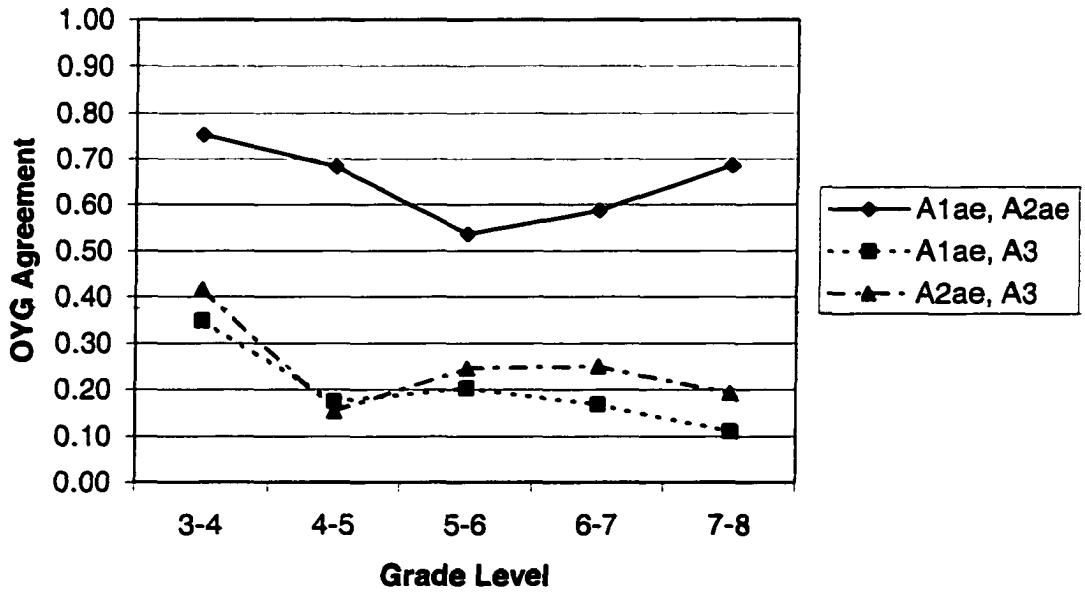

Figure 4.39. Plots of κ of the OYG decision between each pair of Methods $A1_{AE}$, $A2_{AE}$, and A3, by grades in Reading.

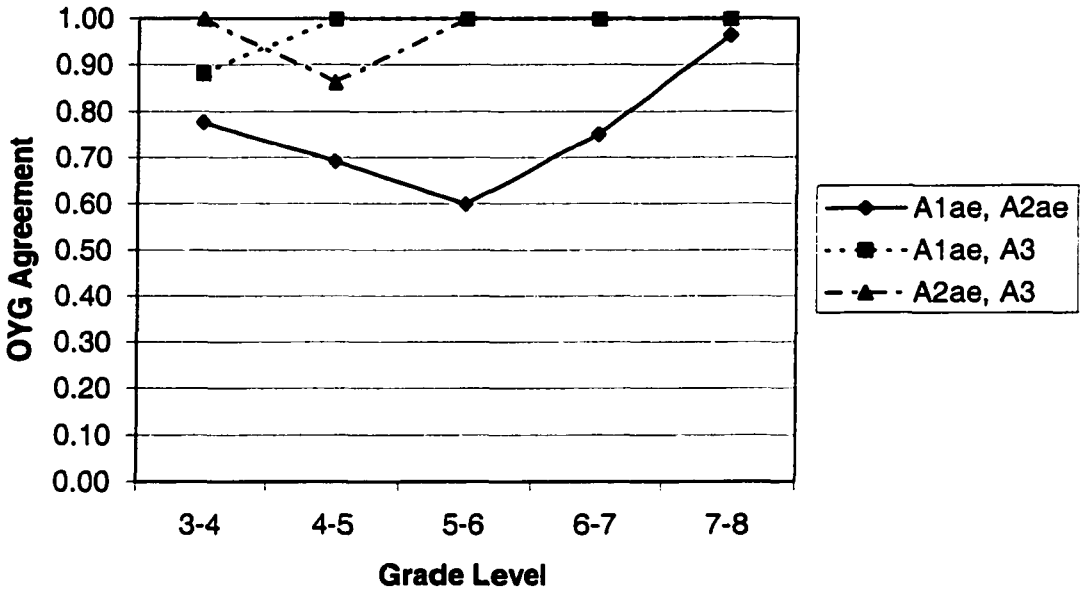

Figure 4.40. Plots of κ / κ_{max} of the OYG decision between each pair of Methods $A1_{AE}$, $A2_{AE}$, and A3, by grades in Reading.

Table 4.43

κ, κ_{max}, and κ / κ_{max} of the OYG Decision between each Pair of Methods $A1_{AE}$, $A2_{AE}$, and A3, by Grades, Mathematics

Grade Level	$A1_{AE}$ and $A2_{AE}$			$A1_{AE}$ and A3			$A2_{AE}$ and A3		
	κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}	κ	κ_{max}	κ / κ_{max}
3-4	.746	1.00	.749	.523	.523	1.00	.523	.523	1.00
4-5	.795	.994	.800	.429	.429	1.00	.412	.426	.966
5-6	.803	.915	.877	.175	.175	1.00	.204	.204	1.00
6-7	.639	.773	.826	.220	.238	.923	.353	.353	1.00
7-8	.706	.811	.869	.142	.142	1.00	.201	.201	1.00

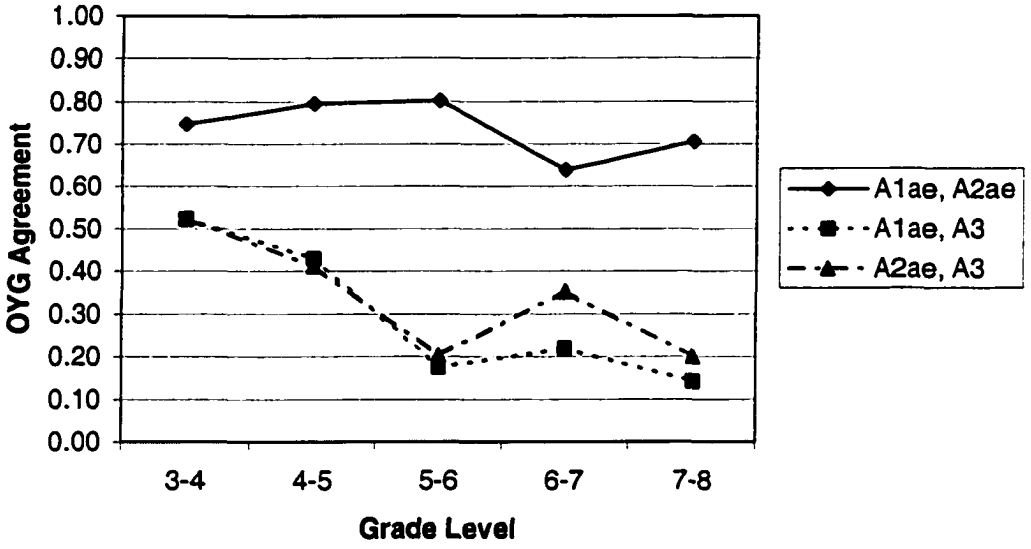

Figure 4.41. Plots of κ of the OYG decision between each pair of Methods $A1_{AE}$, $A2_{AE}$, and A3, by grades in Mathematics.

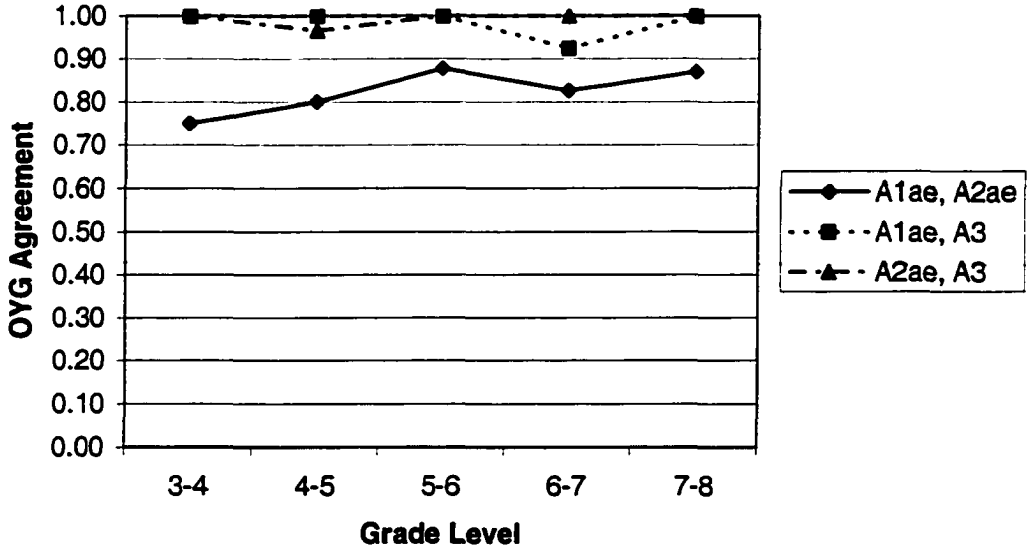

Figure 4.42. Plots of κ / κ_{max} of the OYG decision between each pair of Methods $A1_{AE}$, $A2_{AE}$, and A3, by grades in Mathematics.

The plots of κ in Figures 4.39 and 4.41 show better OYG agreement from Methods $A1_{AE}$ and $A2_{AE}$ than either with Method A3. The agreement (κ) is poor between Method A3 and the other two methods because A3 results in a higher proportion of units achieving OYG than $A1_{AE}$ or $A2_{AE}$. As was demonstrated in earlier analyses, the differences in OYG proportion may result in fewer opportunities for agreement between methods, which forces the maximum possible agreement to decrease. The higher proportion of OYG from Method A3 also explains the large κ / κ_{max} values from Tables 4.42 and 4.43 between A3 and other two methods.

The agreement (κ) between $A1_{AE}$ and $A2_{AE}$ is slightly more consistent across grade levels in Mathematics than it is in Reading. The plots in Figure 4.39 show a decrease in the agreement between $A1_{AE}$ and $A2_{AE}$ at grade levels 5-6 and 6-7, indicating a specific grade level effect in Reading.

Results for Category V:

Minimum Size Criteria for School/Grade Units

The last category had only one research question to determine the number of school/grade units that would be omitted from analyses if different minimum unit size criteria were used. The results from this analysis are listed in Table 4.44 and Figure 4.43.

Table 4.44

Proportion of School/Grade Units Falling Below Minimum Size Criteria

Unit Size	Proportion of Units Below Criteria
5	8.8 %
6	11.0 %
7	12.3 %
8	13.8 %
9	14.9 %
10	15.8 %
11	16.8 %
12	17.8 %
13	18.7 %
14	19.7 %
15	20.6 %
16	21.2 %
17	22.0 %
18	22.7 %
19	23.7 %
20	24.3 %
21	25.0 %
22	25.8 %
23	26.4 %
24	27.1 %
25	27.6 %

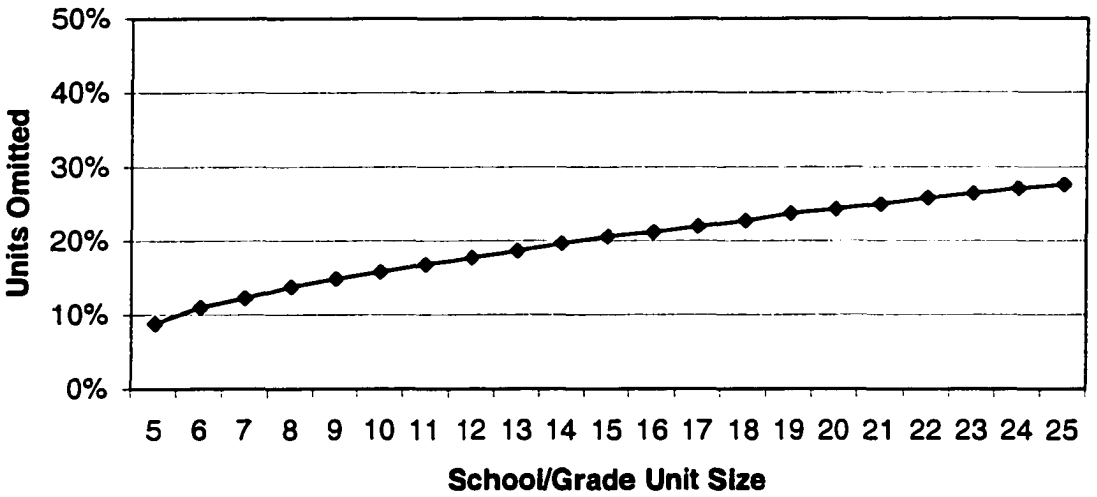

Figure 4.43. Plot of the proportion of school/grade units falling below minimum size criteria.

The data in the Table 4.44 and Figure 4.43 show that the current minimum unit size criterion of eight excludes 13.8 % of potentially eligible school/grade units. If the minimum were 15, approximately 20.6% of eligible units would be omitted from analyses. There is a gradual increase in the proportion of units excluded between unit sizes of 5 to 25. The problem is that for units of size 5 or less, the criterion would have already omitted nearly 9% of the units, accounting for a meaningful proportion of schools that would not receive growth information to serve as feedback.

Another factor relating to unit size is that small units are probably not randomly distributed throughout Arizona. Unit size is likely to differ between rural and urban school districts. The criterion related to unit size may systematically under-represent certain student populations such as Native American students.

CHAPTER 5

SUMMARY AND CONCLUSIONS

The purpose of this study was to critique the method used by Arizona to measure academic progress attributable to schools. In the process, alternative methods using NCE scores (Method A1), PR scores (Method A2), and stanine scores (Method A3) were presented and compared to the scaled score method. A variation of the scaled score method (AZ_{NC}) without the correction for RTM was proposed to examine the effects of the correction. Two variations of the NCE and PR score methods were constructed (Methods $A1_{AE}$ and $A2_{AE}$) with adjusted passing criteria to examine the effect of accounting for measurement error. The goal was to recommend a superior alternative to Arizona's method if one exists. The summary and conclusions are grouped by the five main categories used throughout this current research.

Summary of Category I

The results from this research category suggested a grade-level dependency in Method AZ. While Methods A1 and A2 both reveal differences in the proportions of schools achieving OYG across grade levels, they both tend to reveal the same differences. Method AZ does not show the same pattern of proportions across grade levels and is not consistent with Methods A1 or A2. Method AZ showed a significant reduction in the number of schools achieving OYG during the fifth grade in Reading. This reduction was due likely to the use of incorrect expected growth values, not because of actual student performance. For example, if Method A1 had been used in the initial

MAP report, 50% of the fifth grade schools would have achieved OYG in Reading rather than the reported 26%.

All of the alternative methods showed that the proportion of schools achieving OYG decreased during grades five and seven in both Reading and Mathematics, (although not as drastically as Method AZ did for fifth grade Reading). The causes of the drops are not apparent from results of this study. These fluctuations may represent actual teaching problems in grades five and seven in Arizona or may result because of curricular misfit. The content covered on the Stanford 9 may not match the content taught at these grades in Arizona. To determine if curricular misfit occurred, a content validity study should be undertaken. The drops the proportion of OYG in these two grades may also be attributable to the norming sample in the Stanford 9. When the Stanford 9 was normed, fourth graders may have been lower achieving than fifth graders. If this were the case, then comparing the change of a cohort across two years to a stratified sample from one year may explain the apparent lack of growth during the fifth grade. To determine if the norming sample were the cause, growth in schools from other states could be examined to determine if the patterns across grade levels are consistent.

The results showed that Method A2 yielded consistently higher proportions of schools achieving OYG than Method A1 across all grades in Reading and Mathematics. While this finding appears consistent in this study, higher proportions from A2 may not always be the case. Method A1 computes OYG as a function of average growth and is sensitive to large academic gains or losses made by students. For example, if a teacher made more than one year's worth of progress with a particular student, under Method A1,

this progress could compensate for a student who had not quite made one year's worth of growth. A compensation strategy may sound attractive to teachers; however, a single student making far less than one year's worth of progress can also bring down the mean of the classroom. In the 1998-1999 data used for this research, it would appear that Method A2 yields more favorable results; however it may not always be the case, and would be difficult to predict for future years.

The findings from this category also showed that Methods A1 and A2 are more likely to assign the same OYG decision to each school/grade unit, while Method AZ yields less consistent OYG decisions across grade levels. The lack of consistency in OYG decisions from Method AZ is due likely to the expected growth values used in the method as well as the correction for regression to the mean. The general agreement of the OYG decisions between Methods A1 and A2 suggest that either method can be used without incurring specific grade-level by method effects. Like the OYG decisions, Methods A1 and A2 are also more likely to assign the same Star Rating to each school/grade unit, while Method AZ results in less consistent Star Ratings across grade levels.

This current research predicted that because Method A1 is sensitive to large gains or losses of students, there would be a greater discrepancy between OYG decisions from Methods A1 and A2 in smaller units where the effects of the outliers would be the greatest. The results did not support this expectation, suggesting that the effects of unit size are not a critical factor in recommending Method A1 over Method A2.

Summary of Category II

It was argued in Chapters 1 and 2 that the correction for RTM in Method AZ was not appropriate. This research showed that overall proportions of schools achieving OYG were not affected greatly by the correction. These results, however, were misleading as the impact of the correction was demonstrated in the analyses where schools were grouped by their initial scaled score means. The correction for RTM results in increasing the proportion of school/grade units achieving OYG when their initial status is above the mean. The correction decreases the number of units achieving OYG when their initial status is below the mean. The correction for RTM in Arizona's method resulted in school/grade units with low initial status having to make more than one year's worth of academic growth to achieve OYG. If the correction had not been used, there would have been more growth demonstrated in schools with low initial status.

The classic argument for correcting for RTM is that it makes it easier to demonstrate growth for schools with high initial status. Although this argument is valid, the correction procedures used in the Arizona method (based on the correlation between pre and post measures) were not appropriate. More complex models, such as the HLM model used in the Dallas system, can incorporate the effects of more variables to make adjustments to scores. For example, the initial status of the school/grade unit can be hierarchically modeled. While Arizona could implement an HLM structure it would require the transition from a simple approach of computing growth to a more complex method that may not result in a distinct advantage. In addition, the low scoring schools

would still be forced to achieve more than one year's worth of growth just to maintain OYG.

Summary of Category III

A problem with using simple growth is that it tends to be correlated with initial status. Growth that has been adjusted for initial status or corrected for RTM tends be less correlated with initial status. All of the alternative methods in this research use simple growth. The results tended to confirm correlations between initial status and the growth indicators; but for a majority, the magnitudes of the correlations were not large enough to warrant dismissing the simple growth indicators.

The findings showed that the correction for RTM in Method AZ did not systematically reduce the dependence of initial status and the growth indicators. In fact, for some grade levels, the correlations increased. These inconsistent findings suggest further evidence for not using the standard correction for RTM.

The growth indicators for A1, A2, and A3 exhibited low correlations with initial status, with the exception of the growth that occurred from seventh to eighth grade. This finding was consistent in both Reading and Mathematics. The reasons for the increased correlations for eighth grade were not evident from the results of the analyses. Because all three methods (A1, A2, and A3) showed this increased correlation, it may suggest a grade level effect rather than a method dependency.

Finally, the correlations between the amount of growth used to compute the Star Ratings and initial status were larger than the correlations between the OYG indicators and initial status, suggesting that the OYG indicator is less dependent on initial status

than the Star Rating. However, the OYG and the Star Rating were designed to provide different information, so this finding does not suggest the removal of the Star Rating.

Summary of Category IV

The results from this category showed that the proportion of school/grade units achieving OYG increases when error is accounted for (as in Methods $A1_{AE}$, $A2_{AE}$, or A3). There is a more important practical implication of accounting for error over that of increasing the proportions of OYG. Methods AZ, A1, and A2 do not take error into account. Therefore, they run a greater risk of not identifying OYG when a school/grade unit actually has made adequate growth. The OYG decision for a school carries serious consequences. When the decision is made that a school has not achieved OYG, there needs to be a greater degree of certainty that it is a valid conclusion. Accounting for error allows for greater certainty.

Among Methods $A1_{AE}$, $A2_{AE}$, and A3, Method A3 (with stanines) yielded the highest proportions of school/grade units achieving OYG, particularly in Mathematics, where 91% to 99% of schools achieved OYG. These high proportions of OYG suggest that the use of stanines may be overly conservative for identifying under-performing schools. If however, the intent were to identify those schools that demonstrated the least amount of academic growth, then Method A3 would flag only the lowest 1% to 9%.

An additional benefit to a less sensitive measure of growth is that it may offset the public's and legislative tendency to over-react to small (statistically insignificant) decreases in growth. This is because Method A3 is likely to identify only those schools

that greatly under-perform the other schools in the state. Attention towards these under-performing schools is therefore, more likely to be justified.

Method $A2_{AE}$ is the next most conservative and $A1_{AE}$ is the least conservative, (although none of Methods $A1_{AE}$, $A2_{AE}$, or A3 resulted with fewer than 59% of schools achieving OYG). $A1_{AE}$ and $A2_{AE}$ may be more appropriate if the intent is to identify a larger proportion of schools that need assistance. In this research, the OYG decisions in Methods $A1_{AE}$ and $A2_{AE}$ were adjusted by subtracting one standard error from the criteria. The fact that one standard error was chosen rather than two was an arbitrary choice and could have been adjusted to identify a larger or smaller proportion of schools as not achieving OYG. For example, using one and a half standard errors rather than one, would have identified a smaller proportion of schools as not achieving OYG. Conversely, reducing the standard error factor would lead to an increase in the proportion of schools not achieving OYG. While Method A3 accounts for error, it does not have the same flexibility to adjust the proportions of schools achieving OYG. Criterion adjustments would need to be incorporated into Method A3 to make it more flexible.

Summary of Category V

The stability of the academic growth estimate for each school/grade unit is a function of the variability of student growth and the number of students in the unit. Currently, there is a requirement of at least eight students in order for a unit to be included in the analyses. The problem is that eight observations will not typically provide a stable estimate of true growth. The result is that the growth estimates will fluctuate

more from year to year for school/grade units with fewer students. By increasing the minimum unit size requirement, the variability of the estimates would be decreased.

Unfortunately, recommending that the unit size requirement be increased would overlook another issue. The results of this research showed that in Arizona approximately 14% of eligible school/grade units were omitted because they had fewer than eight students. If the minimum size requirement were increased to fifteen students, approximately 21% of the eligible school/grade units would have been omitted and would not have feedback on their academic efforts. If there were a smaller proportion of eligible units in Arizona with small sizes, then increasing the minimum requirement would be a more feasible recommendation.

Two recommendations that may reduce the impact of the instability associated with small unit size. First, if methods which take error into account were used, such as $A1_{AE}$, $A2_{AE}$ or A3, then the instability of the growth estimates might have a smaller impact than it would have for methods without accounting for error. Second, as growth estimates are accumulated for a particular school/grade unit across multiple reporting years, rather than using each estimate to assess growth, two- or three-year running averages could be computed. Averaged estimates would result in a more stable growth indicator for units with fewer students.

In conclusion, this research has demonstrated that Arizona's initial approach had methodological and performance problems. While it is relatively easy to criticize an existing program, it is more difficult to offer a solution that can function flawlessly under multiple conditions. Arizona's initial method is not recommended for future use;

however, there is no clear "best choice" among the alternative methods presented. As was demonstrated in this research, each alternative has strengths and weaknesses. Using Method A1 would result in a higher proportion of units achieving OYG over Method A2 if a teacher's average NCE gains with students would typically be greater than their average NCE losses. These gains and losses would be difficult to predict in future years, so the strategy of maintaining PR scores (as in Method A2) may be a better choice in the long run. Finally, a method that takes error into account should be implemented. Accounting for error in the OYG decisions allows control over the possibility of misidentification of failing schools as well as the proportion of schools that are identified as needing assistance.

Limitations of the Study

Only alternative methods that used simple (unadjusted) growth were proposed. Other models such as the Dallas HLM model yield growth estimates adjusted to account for differences in external covariates that exist between school/grade units. Models of this sort may well yield fairer, more informative measures regarding school effectiveness. These models were not under consideration in this research and thus the effects are not available for comparison.

While the data used in this study allow for generalization to the State of Arizona, there were findings, such as the decreases in the proportion of units achieving OYG in Mathematics for grade levels 6-7 and 7-8, that cannot be adequately explained. Data from other states are needed to draw conclusions about the State of Arizona relative to national trends.

APPENDIX A

Student-Level Matching Process

The record matching process was conducted by the Arizona Department of Education. The individual student records for two consecutive years of Stanford 9 achievement scores were used. A combination of student identification numbers, last name, first name, date of birth and gender were used to match the pre-measure (1998) with the post-measure (1999) for each student within a school.

Student Inclusion Criteria

Student scores were included in the analyses if four criteria were met.

1. The student took the exam in the same school in both 1998 and 1999.
2. No special testing accommodations were provided to the student.
3. The student did not re-take the same level exam the second year.
4. The student had a valid score in a subject area for both years.

School/Grade Exclusion Criteria

A school/grade level was omitted from analysis if it met any of the following conditions:

1. There were fewer than 8 students in the unit.
2. Less than 25% of eligible students in a grade level were matched across consecutive years.
3. The grade level did not have scores on record for both years.
4. The school did not contain at least two adjoining grade level that are 3 to 8.

APPENDIX B

Table B.1

Range Mean and Standard Deviation of Adjusted Growth in Method AZ by Grade and Star Rating, Reading

Grade Level	Star Rating	N	Minimum	Maximum	Mean	Standard Deviation
3-4	1	128	-6.3068	23.3744	18.2410	4.9950
	2	129	23.5742	27.6271	25.774	1.1690
	3	129	27.6313	30.9666	29.1959	1.0111
	4	129	30.9743	34.6094	32.7336	1.0023
	5	128	34.6104	57.5579	38.4217	4.0754
4-5	1	131	-8.8962	11.0257	8.0211	3.2471
	2	132	11.0369	13.9501	12.5549	0.8797
	3	132	13.9776	16.2388	15.0036	0.6772
	4	132	16.2427	19.0346	17.5817	0.7652
	5	132	19.0545	32.7827	22.1847	2.8496
5-6	1	89	-29.1666	8.8138	5.1949	4.9631
	2	89	8.9197	11.6648	10.3397	0.8398
	3	89	11.6789	13.8423	12.6562	0.6049
	4	89	13.9816	16.1590	14.9760	0.6713
	5	89	16.1626	45.9419	19.6617	4.5657
6-7	1	51	-5.0261	12.5557	8.6940	3.7209
	2	51	12.7549	16.4897	14.8883	0.9873
	3	52	16.7147	19.9483	18.5580	0.8855
	4	51	19.9861	23.9088	21.8154	1.1812
	5	51	23.9763	34.5635	26.5337	2.2981
7-8	1	66	-4.7937	10.3996	7.76049	2.9963
	2	67	10.4079	12.3233	11.4182	0.5238
	3	66	12.3434	14.1167	13.2619	0.4656
	4	67	14.1428	16.4353	15.1584	0.6657
	5	128	-6.3068	23.3744	18.2410	4.9950

APPENDIX B - *Continued*

Table B.2

Range Mean and Standard Deviation of Adjusted Growth in Method AZ by Grade and Star Rating, Mathematics

Grade Level	Star Rating	N	Minimum	Maximum	Mean	Standard Deviation
3-4	1	129	-18.6947	24.9425	19.3589	6.7469
	2	130	24.9645	30.6877	27.8220	1.7043
	3	130	30.7066	34.7913	32.7619	1.1887
	4	130	34.9466	40.0388	37.4575	1.5984
	5	129	40.0918	68.5401	45.4311	4.7622
4-5	1	132	-1.7977	17.4052	13.2093	3.8595
	2	132	17.4271	22.6316	20.1120	1.5202
	3	133	22.6417	26.9788	24.7391	1.3121
	4	132	27.0031	31.4630	29.0405	1.3615
	5	132	31.5082	59.2942	36.6947	5.0558
5-6	1	89	-8.1692	13.8038	9.2793	3.5321
	2	90	13.8426	19.3268	16.7381	1.4998
	3	90	19.3659	23.4953	21.5350	1.2018
	4	90	23.5328	27.5793	25.4372	1.2172
	5	90	27.6339	62.1327	32.7274	5.8782
6-7	1	51	-13.6118	9.1862	4.9179	3.8595
	2	51	9.1968	11.8297	10.5713	.8320
	3	52	11.8379	14.9241	13.2500	.9275
	4	51	15.1249	19.3123	17.2967	1.2416
	5	51	19.4418	46.7048	25.4333	5.7374
7-8	1	66	-9.6933	5.8131	2.9816	2.8636
	2	67	5.8341	8.7156	7.4707	.8556
	3	66	8.8402	10.7805	9.7339	.5688
	4	67	10.8300	14.9405	12.5939	1.2185
	5	66	15.0112	39.9635	19.3612	4.4836

APPENDIX B - *Continued*

Table B.3

Range Mean and Standard Deviation of Simple Growth in Method A1 by Grade and Star Rating, Reading

Grade Level	Star Rating	N	Minimum	Maximum	Mean	Standard Deviation
3-4	1	128	-14.7120	1.0764	-1.4642	2.5234
	2	129	1.0907	2.7866	2.0495	.4948
	3	129	2.8047	4.3120	3.5827	.4380
	4	129	4.3146	5.9618	5.0426	.4558
	5	128	5.9854	19.2212	8.1152	2.2490
4-5	1	131	-12.4421	-2.2923	-3.9083	1.6216
	2	132	-2.2903	-.7025	-1.4070	.4716
	3	132	-.6965	.6319	-.0493	.3722
	4	132	.6333	2.3346	1.3923	.4773
	5	132	2.3440	10.4440	4.0937	1.6882
5-6	1	89	-23.4654	-.0469	-2.0719	2.8992
	2	89	-.0205	1.4421	.7235	.4479
	3	89	1.4562	2.6528	1.9971	.3632
	4	89	2.6667	4.2261	3.3776	.3973
	5	89	4.2429	22.4500	6.6215	2.8856
6-7	1	51	-11.8432	-1.7222	-3.5249	2.0654
	2	51	-1.6577	-.1882	-.9022	.4552
	3	52	-.1555	1.1333	.5087	.3742
	4	51	1.1383	2.9018	1.9758	.5155
	5	51	2.9023	8.2900	4.1247	1.2580
7-8	1	66	-10.1000	-.5179	-2.2060	1.7974
	2	67	-.5160	.7926	.1400	.3728
	3	66	.8155	1.9554	1.4250	.3409
	4	67	1.9615	3.7722	2.7618	.5400
	5	66	3.7885	12.7500	5.5417	1.6475

APPENDIX B - *Continued*

Table B.4

Range Mean and Standard Deviation of Simple Growth in Method A1 by Grade and Star Rating, Mathematics

Grade Level	Star Rating	N	Minimum	Maximum	Mean	Standard Deviation
3-4	1	129	-27.5727	-.3174	-3.5512	3.7983
	2	130	-.3122	2.4473	1.1926	.8065
	3	130	2.4656	4.8763	3.5345	.7124
	4	130	4.9012	7.8322	6.2076	.8902
	5	129	7.8861	25.9529	10.9739	2.8428
4-5	1	132	-15.3333	-1.9446	-4.5350	2.4583
	2	132	-1.8904	.7794	-.4160	.7961
	3	133	.7816	2.9262	1.8297	.5952
	4	132	2.9404	5.7631	4.1661	.8093
	5	132	5.7695	23.3333	8.5794	2.9625
5-6	1	89	-8.0088	1.3661	-.7512	1.8760
	2	90	1.3889	4.3401	2.8828	.8272
	3	90	4.3493	6.2528	5.2671	.5761
	4	90	6.2562	8.4915	7.2779	.6773
	5	90	8.5382	29.8235	11.3773	3.1957
6-7	1	51	-16.0256	-2.8238	-5.2449	2.2682
	2	51	-2.8150	-1.3412	-2.0501	.4512
	3	52	-1.3301	.5986	-.4055	.5925
	4	51	.7562	2.6831	1.7384	.6297
	5	51	2.8000	18.7526	5.7415	3.2718
7-8	1	66	-10.8000	-1.7619	-3.4308	1.7466
	2	67	-1.7569	-.0552	-.8485	.5131
	3	66	-.0470	1.0575	.4821	.3423
	4	67	1.0612	2.6711	1.8401	.4988
	5	66	2.6921	16.0571	5.3160	2.6652

APPENDIX B - *Continued*

Table B.5

Range Mean and Standard Deviation of the Proportion of Students Achieving OYG in Method A2 by Grade and Star Rating, Reading

Grade Level	Star Rating	N	Minimum	Maximum	Mean	Standard Deviation
3-4	1	126	.0400	.5536	.4604	.0984
	2	131	.5556	.6264	.5926	.0205
	3	129	.6269	.6829	.6560	.0171
	4	129	.6832	.7451	.7124	.0181
	5	128	.7458	1.0000	.8078	.0521
4-5	1	133	.0000	.4211	.3582	.0610
	2	131	.4217	.4848	.4560	.0182
	3	131	.4853	.5435	.5142	.0174
	4	132	.5439	.6190	.5771	.0214
	5	132	.6197	.9211	.6855	.0688
5-6	1	89	.1250	.5238	.4495	.0703
	2	88	.5246	.5859	.5571	.0186
	3	90	.5870	.6336	.6123	.0134
	4	89	.6338	.6892	.6592	.0154
	5	89	.6900	.9412	.7695	.0678
6-7	1	50	.1250	.4556	.3749	.0741
	2	52	.4583	.5161	.4878	.0177
	3	52	.5172	.5735	.5451	.0170
	4	50	.5755	.6324	.6044	.0166
	5	52	.6364	.8462	.6987	.0560
7-8	1	69	.1250	.5000	.4349	.0720
	2	65	.5060	.5556	.5321	.0155
	3	65	.5566	.6034	.5792	.0132
	4	67	.6051	.6721	.6349	.0209
	5	66	.6748	.9000	.7419	.0531

APPENDIX B - *Continued*

Table B.6

Range Mean and Standard Deviation of the Proportion of Students Achieving OYG in Method A2 by Grade and Star Rating, Mathematics

Grade Level	Star Rating	N	Minimum	Maximum	Mean	Standard Deviation
3-4	1	129	.0000	.5066	.4063	.1032
	2	134	.5068	.6000	.5585	.0280
	3	125	.6029	.6809	.6439	.0214
	4	129	.6818	.7755	.7241	.0262
	5	131	.7778	1.0000	.8498	.0549
4-5	1	131	.1111	.4483	.3560	.0788
	2	132	.4500	.5410	.4974	.0264
	3	133	.5417	.6263	.5859	.0244
	4	133	.6275	.7167	.6678	.0265
	5	132	.7188	1.0000	.8017	.0653
5-6	1	89	.2059	.5833	.4907	.0759
	2	91	.5867	.6857	.6382	.0301
	3	92	.6863	.7500	.7233	.0201
	4	88	.7534	.8267	.7861	.0222
	5	89	.8272	1.0000	.8878	.0500
6-7	1	51	.1282	.3929	.3126	.0687
	2	51	.3946	.4686	.4372	.0231
	3	52	.4706	.5556	.5151	.0266
	4	51	.5560	.6571	.6018	.0277
	5	51	.6582	1.0000	.7480	.0855
7-8	1	66	.0000	.4400	.3607	.0793
	2	67	.4407	.5036	.4778	.0194
	3	66	.5063	.5620	.5327	.0162
	4	67	.5625	.6512	.6066	.0256
	5	66	.6522	1.0000	.7457	.0873

APPENDIX B - *Continued*

Table B.7

Range Mean and Standard Deviation of the Proportion of Students Achieving OYG in Method A3 by Grade and Star Rating, Reading

Grade Level	Star Rating	N	Minimum	Maximum	Mean	Standard Deviation
3-4	1	128	.2000	.7222	.6320	.0924
	2	129	.7231	.7755	.7522	.0162
	3	129	.7759	.8235	.7969	.0132
	4	127	.8243	.8661	.8437	.0123
	5	130	.8667	1.0000	.9136	.0382
4-5	1	131	.2222	.6176	.5474	.0677
	2	132	.6190	.6765	.6493	.0173
	3	132	.6769	.7292	.7043	.0152
	4	132	.7300	.7813	.7550	.0145
	5	132	.7818	1.0000	.8398	.0498
5-6	1	90	.3333	.7051	.6441	.0700
	2	87	.7059	.7664	.7363	.0169
	3	90	.7667	.8043	.7847	.0108
	4	89	.8061	.8539	.8280	.0131
	5	89	.8542	1.0000	.8981	.0396
6-7	1	51	.2432	.6275	.5585	.0780
	2	52	.6307	.7000	.6643	.0230
	3	51	.7018	.7368	.7193	.0102
	4	51	.7381	.7870	.7593	.0145
	5	51	.7895	1.0000	.8357	.0425
7-8	1	68	.4444	.6923	.6307	.0638
	2	65	.6933	.7429	.7184	.0142
	3	66	.7437	.7795	.7606	.0111
	4	67	.7809	.8333	.8040	.0149
	5	66	.8361	1.0000	.8905	.0450

APPENDIX B - *Continued*

Table B.8

Range Mean and Standard Deviation of the Proportion of Students Achieving OYG in Method A3 by Grade and Star Rating, Mathematics

Grade Level	Star Rating	N	Minimum	Maximum	Mean	Standard Deviation
3-4	1	130	.0909	.6774	.5763	.1062
	2	130	.6780	.7544	.7214	.0224
	3	129	.7547	.8182	.7839	.0179
	4	130	.8205	.8862	.8539	.0196
	5	129	.8873	1.0000	.9312	.0346
4-5	1	132	.2222	.6122	.5151	.0856
	2	132	.6134	.6970	.6581	.0241
	3	133	.6984	.7590	.7295	.0170
	4	135	.7595	.8333	.7979	.0219
	5	129	.8356	1.0000	.8979	.0488
5-6	1	88	.3824	.7436	.6626	.0683
	2	91	.7442	.8085	.7772	.0196
	3	90	.8088	.8667	.8376	.0180
	4	90	.8675	.9167	.8885	.0150
	5	90	.9175	1.0000	.9552	.0297
6-7	1	51	.3077	.6078	.5104	.0768
	2	51	.6111	.6667	.6399	.0164
	3	52	.6750	.7391	.7088	.0191
	4	52	.7456	.8136	.7747	.0212
	5	50	.8163	1.0000	.8822	.0606
7-8	1	66	.1538	.6301	.5483	.0866
	2	67	.6305	.6885	.6608	.0172
	3	66	.6892	.7342	.7136	.0138
	4	66	.7364	.7955	.7651	.0181
	5	67	.8000	1.0000	.8555	.0502

APPENDIX B - *Continued*

Table B.9

Range Mean and Standard Deviation of Simple Growth in Method AZ_{NC} by Grade and Star Rating, Reading

Grade Level	Star Rating	N	Minimum	Maximum	Mean	Standard Deviation
3-4	1	128	-8.8000	23.6429	18.8032	5.3174
	2	130	23.6512	27.5000	25.6691	1.0444
	3	128	27.5429	30.5181	29.0429	.8471
	4	129	30.5636	34.1282	32.0884	1.1115
	5	128	34.1818	61.5455	38.7968	4.8032
4-5	1	131	-7.5000	9.9167	6.7701	2.9466
	2	132	9.9211	13.7237	11.9373	1.0619
	3	132	13.7292	16.3256	15.0249	.8087
	4	132	16.3667	19.8276	17.9645	.9542
	5	132	19.8306	36.0000	23.6381	3.4327
5-6	1	89	-33.6923	8.1351	4.6480	5.3118
	2	89	8.1389	11.0448	9.6822	.8279
	3	89	11.0488	13.3000	12.1876	.6775
	4	89	13.3538	16.4038	14.9060	.9222
	5	89	16.4085	51.0714	21.4030	5.4491
6-7	1	51	-5.0000	13.5882	10.5107	3.8154
	2	51	13.6277	16.7789	15.4259	.9199
	3	52	16.8100	19.7480	18.2785	.8865
	4	51	19.8889	22.4512	21.1521	.8255
	5	51	22.5000	32.6333	25.1243	2.3325
7-8	1	66	-7.7143	9.6480	6.4887	3.3907
	2	67	9.6613	12.1122	10.8886	.6753
	3	66	12.1216	14.2128	13.2951	.6580
	4	67	14.2179	17.7000	15.8031	1.0551
	5	66	17.7037	34.5000	21.1846	13.3538

APPENDIX B - *Continued*

Table B.10

Range Mean and Standard Deviation of Simple Growth in Method AZ_{NC} by Grade and Star Rating, Mathematics

Grade Level	Star Rating	N	Minimum	Maximum	Mean	Standard Deviation
3-4	1	129	-27.0000	24.3438	18.4164	6.9543
	2	130	24.4667	30.2714	27.6575	1.6896
	3	130	30.3333	34.8750	32.4248	1.3770
	4	130	35.0606	40.5789	37.6060	1.5700
	5	129	40.6471	74.5294	46.7333	5.3658
4-5	1	132	-9.2222	17.5167	12.9355	4.4677
	2	132	17.5385	22.7778	20.3314	1.5557
	3	133	22.7959	26.5385	24.6174	1.0209
	4	132	26.5524	31.3065	28.7400	1.4181
	5	132	31.3600	62.6667	37.1696	5.5702
5-6	1	89	-5.0000	14.4211	9.9465	3.6624
	2	90	14.4954	19.0597	16.9640	1.3459
	3	90	19.1667	22.8824	21.2472	1.0493
	4	90	22.9524	27.1500	25.0523	1.2075
	5	90	27.1613	65.9412	32.5146	6.1252
6-7	1	51	-14.4615	8.3077	4.3158	4.0972
	2	51	8.4074	12.3544	10.3519	1.1123
	3	52	12.5229	15.2727	13.9461	.8427
	4	51	15.3137	19.7120	17.6481	1.3464
	5	51	19.8182	48.2105	25.1913	5.8476
7-8	1	66	-11.7692	5.8614	2.9629	3.2594
	2	67	5.8690	8.9366	7.6294	.8589
	3	66	8.9474	10.8831	9.8779	.5881
	4	67	10.9096	14.3313	12.5875	1.0081
	5	66	14.3662	39.4286	19.0828	4.8665

REFERENCES

Airasian, P. W. (1997). Oregon teacher work sample methodology: Potential and problems. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation method?* (pp. 46-52). Thousand Oaks, CA: Corwin Press.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Bishop, M M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis.* Cambridge, MA: MIT Press.

Garcia, D., & Aportela, A. (2000). *Arizona measure of academic progress: A first look at growth in Arizona schools technical document.* A report for Lisa Graham Keegan, Superintendent of Public Instruction, Arizona Department of Education.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts.* New York: Guilford Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and* Psychological Measurement, *1*, 37-46.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation design & analysis issues for field settings.* Boston: Houghton Mifflin.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23*, 283-298.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.). (pp. 105-146). Phoenix, AZ: Oryx Press.

Haberman, S. J. (1979). *Analysis of qualitative data: Volume 1 introductory topics.* New York: Academic Press.

Harcourt Brace. (1997). *Stanford achievement test series (9th Ed.): Technical data report.* San Antonio, TX.

Kennedy, J. J. (1983). *Analyzing qualitative data: Introductory log-linear analysis for behavioral research.* New York: Praeger.

Kerbow, D. W. (1996). *Pervasive student mobility: A moving target for school improvement.* Chicago: University of Chicago, Center for School Improvement.

Lord, F. M. (1980). *Practical applications of item response theory*. Hillsdale, NJ: Earlbaum.

Linn, R. (1981). Determining pretest-posttest performance changes. In R. A. Berk (Ed.), *Educational and evaluation methodology: The state of the art* (pp. 85-109). Baltimore: John Hopkins University Press.

Lund, T. (1989). The statistical regression phenomenon: I a metamodel. *Scandinavian Journal of Psychology, 30*, 2-11.

Lund, T. (1989). The statistical regression phenomenon: II application of a metamodel. *Scandinavian Journal of Psychology, 30*, 12-29.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scales, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.). (pp. 475-483). Phoenix, AZ: Oryx Press.

Proposition 301. Amendment to Arizona Revised Statute A.R.S. §12-241. Passed in November of 2000.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation method?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Schalock, H. D., Schalock, M., & Girod, G. (1997). Teacher work sample methodology used at Western Oregon State College. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation method?* (pp. 62-73). Thousand Oaks, CA: Corwin Press.

Slinde, J.A., & Linn, R.L. (1979). A note on vertical equating via the rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Psychology, 16*, 159-165.

Sykes, G. (1997). On Trial: The Dallas value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation method?* (pp. 110-119). Thousand Oaks, CA: Corwin Press.

Thum, Y. M. & Bryk, A. S. (1997). Value-added productivity indicators: The Dallas system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation method?* (pp. 100-109). Thousand Oaks, CA: Corwin Press.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications.* Thousand Oaks, CA: Sage.

Upton, G. J. (1978). *The analysis of cross-tabulated data.* New York: Wiley.

Walberg, H. J. & Paik, S. J. (1997). Assessment requires incentives to add value: A review of the Tennessee value-added assessment system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation method?* (pp. 169-178). Thousand Oaks, CA: Corwin Press.

Webster, W. J. (2000, July). *A value-added application of hierarchical linear modeling to the estimation of school and teacher effect.* Paper presented at the First Annual Assessment Conference of the Council of Great City Schools, Portland, OR.

Webster, W. J., Mendro, R. L., Orsak, T. H., & Weerasinghe, D. (1996, April). *The applicability of selected regression and hierarchical linear models to the estimation of school and teacher effects.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.