

2006

## Measuring inter-rater agreement for a school psychology case study rubric

Tracy Kay Spires  
*University of Dayton*

Follow this and additional works at: [https://ecommons.udayton.edu/graduate\\_theses](https://ecommons.udayton.edu/graduate_theses)

---

### Recommended Citation

Spires, Tracy Kay, "Measuring inter-rater agreement for a school psychology case study rubric" (2006).  
*Graduate Theses and Dissertations*. 5733.  
[https://ecommons.udayton.edu/graduate\\_theses/5733](https://ecommons.udayton.edu/graduate_theses/5733)

This Thesis is brought to you for free and open access by the Theses and Dissertations at eCommons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of eCommons. For more information, please contact [mschlangen1@udayton.edu](mailto:mschlangen1@udayton.edu), [ecommons@udayton.edu](mailto:ecommons@udayton.edu).

**MEASURING INTER-RATER AGREEMENT  
FOR A SCHOOL PSYCHOLOGY  
CASE STUDY RUBRIC**

Thesis

Submitted to

The School of Education and Allied Professions of the  
UNIVERSITY OF DAYTON

In Partial Fulfillment of the Requirements for

The Degree

Educational Specialist of School Psychology in Counselor Education and Human  
Services

by

Tracy Kay Spires

UNIVERSITY OF DAYTON

June 2006

SCHOOL PSYCHOLOGY PROGRAM  
DEPARTMENT OF COUNSELOR EDUCATION AND HUMAN SERVICES  
SCHOOL OF EDUCATION AND ALLIED PROFESSIONS

WE HEREBY APPROVE THE THESIS SUBMITTED

BY

TRACY SPIRES

ENTITLED:

Measuring Inter-Rater Agreement for a School Psychology Case Study Rubric

---

AS PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Educational Specialist in School Psychology

[Redacted Signature]

Chair

Date

[Redacted Signature]

Member

Date

[Redacted Signature]

Member

Date

## Abstract

Higher education academic programs are obligated to develop student-learning outcomes and objectively assess student performance. Nationally approved school psychology training programs abide by the National Association of School Psychologists' (NASP) requirements. Trainees are required to participate in field experiences and complete case studies. Although studies have investigated rubric effectiveness and experimental methodologies to evaluate rubric reliability there has been little research on assessment instruments for school psychology case studies. Participants in this study were school psychology graduate students who had a semester long course on conducting case studies. Their ratings of case studies were determined to be accurate or inaccurate and category frequency counts were conducted to determine inter-rater agreement. The results suggest that this rubric can provide a reliable, objective measure of performance on case studies.

## Introduction

Recent education acts and reforms have stressed the importance of school accountability for students' academic progress. These mandated educational acts and reforms have primarily targeted elementary and secondary education programs. Institutions of higher learning are also responding to school accountability issues by implementing curriculum change and developing new objective academic assessments to measure student performance. Accordingly, regional and national credentialing, approval and accrediting agencies set certain criteria that are required for programs at the higher education level. Higher education academic programs are required to develop assessment strategies to measure student performance on the designated learning outcomes (Choinski, Mark, & Murphey, 2003).

The National Association of School Psychologists (NASP) has responded to the need for increased accountability by providing a framework of standards to guide school psychology training programs. The NASP has identified eleven key domains which include: data-based decision-making and accountability; consultation and collaboration; effective instruction and development of cognitive/academic skills; socialization and development of life skills; student diversity in development of learning; school and systems organization, policy development, and climate; prevention, crisis intervention, and mental health; home/school/community collaboration; research and program evaluation; school psychology practice and development; and information technology (NASP, 2000). Program approval by NASP requires that school psychologist trainees demonstrate to their faculty the knowledge and skills in the eleven domains that result in measurable, positive changes in the students that they serve.

The completion of a case study is usually a requirement for trainees in NASP approved school psychology training programs. However, the format for conducting and evaluating case studies in approved programs is determined by each program. Case studies based upon the problem-solving model and incorporating research-based academic and/or behavioral interventions serve to demonstrate knowledge and skills in multiple domains. The current goal is to establish inter-rater agreement of an analytic rubric developed to assess performance on the case studies completed by graduate students in a NASP-approved school psychology training program

#### Literature Review.

##### *Rubrics*

Rubrics are guidelines for scoring student products that provide specific definitions corresponding to a list of requirements that are related to a leveled point system (Mabry, 1999). Rubrics provide an objective method to assess student products. The utilization of rubrics with specific definitions, requirements lists, and point systems increase interrater reliability (Mabry, 1999).

Rubrics tend to be classified into two categories: holistic or analytic. Holistic scoring rubrics tend to view the whole product as more important than the specific parts of the product (Klein et al., 1998). In holistic rubrics, one score is awarded for the quality of the overall product. Conversely, analytic scoring rubrics focus on specific requirements that make up the entire product. Raters that utilize an analytic rubric provide scores on a designated number of parts included in the overall product (Klein et al., 1998).

Klein, et al. (1998) indicate that theoretically, holistic scoring techniques are more subjective with analytic rubrics being more of an objective assessment method. Analytic rubrics provide specific definitions for the scoring of each part of a product; therefore, there is a decreased likelihood of rater bias and extraneous variable impact (Klein et al., 1998).

### *Analytic Versus Holistic Rubrics*

Research has been conducted to determine which scoring rubric, analytic or holistic, is the more reliable assessment technique. Crehan and Hudson (2001) investigated whether statistically significant differences existed between two holistic scoring strategies. The first holistic scoring strategy consisted of one overall score for student written work. The second holistic strategy included score categories with suggested borders between categories (Crehan & Hudson, 2001). Borders helped provide additional information for scoring purposes.

Twenty raters scored two hundred, fifth grade written responses. Half of the raters used the first holistic scoring strategy and the other half of the raters used the holistic strategy that included borders between score categories. The results indicated that there was no statistically significant difference between the two holistic scoring strategies (Crehan & Hudson, 2001). The researchers suggest that both holistic rubrics had similar generalizability and percentage of interrater agreement.

Another study examined whether the utilization of rating augmentation would increase interrater reliability of an analytic scoring rubric (Penny, Johnson, & Gordon, 2000a). Rating augmentation was accomplished when the raters added a "+" to the score if the product met a given benchmark and had additional qualities, but did not meet the

criteria for the next benchmark. Conversely, the raters add a “-“ to the score if the product met a given benchmark with only the minimal requirements for the benchmark, yet exceeded the requirements of the lower benchmark. In this study, two raters scored 120 randomly selected eleventh grade-student writing samples based on a 4-point analytic rubric that assessed the four areas of content/organization, style, conventions of written language, and sentence formation (Penny et al, 2000a). The results revealed that the employment of rating augmentation increased overall interrater reliability. But at the same time the percentage of exact agreement between the raters decreased due to an increase in rating options.

These same authors followed up with an investigation of a holistic rating augmentation scoring rubric to determine if it would increase interrater reliability (Penny, Johnson, & Gordon, 2000b). The researchers defined augmentation exactly as it had been in their previous study with the analytic rubric. Two raters scored 120 randomly selected fifth grade-writing samples based on a 6-point holistic rubric. The holistic rubric was based on the following overall scores: emerging writer, developing writer, focusing writer, experimenting writer, engaging writer, and extending writer. Overall, Penny et al, (2000b) found that using a rating augmentation system increased interrater reliability of the holistic rubric scoring method.

Interrater reliability comparisons between holistic and analytic scoring rubrics were examined using family literacy portfolios developed by Even Start Staff (Johnson, Fisher, Willeke, & McDaniel, 2003). The analytic rubric was comprised of a 5-point system with descriptions at the midpoint (‘3’-evidence of several activities of parent teaching child) and extremes (‘1’- minimal evidence of parent teaching child and ‘5’-



evidence that parent often teaches child). The holistic rubric was based on four overall proficiency levels including, 'Proficient' (family literacy skills established), 'Developing' (some family literacy skills developed), 'Emerging' (few family literacy skills), and 'Not Yet' (no family literacy skills) (Johnson et al., 2003). Across 42 family literacy portfolios, the interrater reliability for the analytic scoring rubric was .74 and for the holistic scoring rubric was .79. It was concluded that the interrater reliability of the mean scores for the family literacy portfolios only reached the minimally acceptable standard for reporting (Johnson et al., 2003).

Research conducted by Klein et al. (1998) also investigated whether there were statistically significant differences in interrater reliability between a holistic and an analytic scoring system. Pairs of readers scored Grade 5 and Grade 8 students' written work twice with the analytic system and twice by the holistic system. In sum, each student's written work was scored a total of four times by different scorers. The analytic rubric consisted of scores ranging from '1' (non-responsive) to '6' (outstanding). A scoring guide provided examples of student responses that corresponded to each score level. The holistic scoring system consisted of one overall score for each part of the student's written work. The holistic method ranged from '1' (worst response to the part) to '5' (best response of the part). Overall, the results indicated that there was no statistically significant difference between the mean scores from the holistic and the analytic rubric scoring systems. However, the results identified a higher reliability coefficient among raters for the analytic scoring method than the holistic scoring method.

*Utilizing Rubrics for Performance Measures*

Schools and educators are finding a continued need for objective assessments to measure performance levels. The strength of the rubric is in the generalizability and flexibility of the instrument, which can measure a wide variety of skills through procedures than can be demonstrated to be technically reliable.

*Current Research*

Recent research in differing education domains has investigated the usefulness of rubrics to objectively assess performance levels. For example, in order to assess the performance level on student problem solving abilities, researchers developed an analytic scoring rubric (Anderson & Puckett, 2003). They identified the problem solving areas of written work, group projects, online discussions, classroom presentation, and portfolios for assessment. It was concluded that traditional tests would not have adequately assessed the range of student problem solving abilities, in the variety of settings which were easily measured by the analytic rubrics (Anderson & Puckett, 2003).

In a related line of research, Choinski et al. (2003) investigated student progress in a library 'for credit' information resource class. Numerous measurement techniques of student progress were evaluated to determine an objective and effective assessment method, including student surveys, pre/post tests, teaching portfolios, skill tests, and grading rubrics. Student surveys were not considered objective measures and typically correlated with student satisfaction in the course rather than student performance level. Pre/post test measurements were considered to be more objective in assessment, yet much more time consuming, in that the measure had to be administered twice. Additionally, the procedure had to differ from the measurement that assessed the course

grade. Teaching portfolios (that would have included surveys, self-evaluations, and peer evaluations) were evaluated; however, the researchers noted that this method would not effectively assess student-learning outcomes. Skill tests were considered to be too time consuming to adequately develop, administer in class, and they also decreased overall instructional time. Rubrics were considered by the researchers to be the preferred method for measuring student learning outcomes because they were most objective and time efficient (Choinski et al., 2003).

Medical school graduates' written communication skills were assessed by using a holistic rubric to determine their level of writing skills on patient notes (Boulet, Rebbecchi, Denton, McKinley, & Whelan, 2004). A holistic rubric (2004) was utilized because patient notes could be scored on medical content, legibility, logic, interpretability, and processes. Overall, Boulet et al. (2004) found that the generalizability values were acceptable for the written communication performance assessment.

#### *Establishing Rubric Validity and Reliability*

Prior to the utilization of a rubric to assess performance level, research must be conducted to verify the reliability and validity of the rubric as a scoring instrument. Previous research denotes numerous experimental methods to examine rubric reliability and validity. Additionally, the research conducted to validate rubric reliability and validity explored rubric implementation for many differing domain areas.

Novak, Herman, and Gearhart (1996) conducted a research study to determine the reliability and validity of a newly developed rubric designed to assess writing skills. Five raters scored 52 elementary school student written work samples using two different

holistic rubrics for each of the following areas: direct assessment, samples of classroom narratives, and narrative collections. The validity of the rubric was determined by comparing the narrative collections with direct assessments and samples of classroom narratives.

The holistic rubric was based on a 6-point system that ranged from '1' (no reference to time, place, action, or conflict) to '6' (fully developed narrative with characters, setting, plot, using extensive language and vocabulary). Each point reference provided a specific definition of requirements for the narrative. Additionally, the researchers compared the newly developed rubric with an already established holistic rubric. The established rubric ranged from '1' (insufficient writer) to '6' (exceptional writer). Again, each score provided a definition of what was required to obtain points at that given benchmark.

Overall, results indicated that greater interrater reliability was evidenced from data collected utilizing the new rubric over the established rubric (Novak et al., 1996). The researchers also found that the new rubric provided more consistent scores across the three domain areas. In contrast, the established rubric scores significantly varied across the three domains. However, the researchers suggest that the results were unable to provide strong statistical evidence due to a small sample size. The evidence found for concurrent validity was varied. Student scores were consistent across all three domains that supported validity of the new rubric and the established rubric. Additionally, student scores increased across grade levels as expected. This provided strong results for the validity of the new rubric. However, strong relationships were found between the new rubric narrative collection scores with each of the comparison measures. Further

analyses suggested that with designated cut points, reliable conclusions could be made for mastery or non-mastery of narrative writing skills utilizing the new rubric, yet not for the comparison collection scores (Novak et al., 1996).

Baker, Abedi, Linn, and Niemi (1995) also conducted research to establish reliability and validity of a rubric scoring system to assess students' performance in an eleventh grade history course. Four raters scored 69 student writing assignments based on the scoring rubric that consisted of six domains: general content quality, prior knowledge, principles/concepts, proportion of text detail, misconception, and argumentation. Each domain was scored according to a 5-point system with ' being "no response" to 5 indicating "highest level of response".

The researchers analyzed the rater's scores across all 6 domains to assess interrater reliability of the rubric scores. They found reliabilities ranging from .84 to .91 for five of the six domain areas. The lowest reliability coefficient was obtained in the domain of misconception, which ranged from .52 to .73. Overall, the researchers concluded that the results established a high level of interrater reliability for the rubric in five of the six domain areas (Baker et al., 1995).

Interrater reliability, validity, and generalization of a History Explanation Rubric for multiple criteria were investigated by assessing eleventh grade students' level of performance in three topic areas: Civil War, Chinese Immigration, and General Immigration (Abedi & Baker, 1995). The History Explanation Rubric consisted of the same six domains as the rubric in their initial study. Four raters scored 68 history essays on each of the six domains. Overall results confirmed high measurement quality for the scoring rubric with high interrater reliability and high generalizability. Their results were

consistent with the results of their previous study. However, it should be noted, that the follow-up research produced more significant results and improved interrater reliability rates.

Research to assess reliability of a scoring rubric and the impact of implementing training for rubric scoring on interrater reliability was conducted by Stuhlmann, Daniel, Dellinger, Denny, and Powers (1999). Initially, the investigation began with a pilot study. The researchers wanted to determine if training was required to understand the scoring rubric or whether the rubric was reliable without specific training. Seven educators scored 34 writing portfolios. The rubric contained six subscales for writing development. The rubric subscales were pre-writing/picture, story, sentence, mechanics/spelling, punctuation and capitalization, and format. Each subscale was based on a 4-point system that ranged from 'no attempt or evidence' to 'proficient or well-developed' (Stuhlmann et al., 1999).

The pilot study results indicate that the interrater reliability ranged from .60 to .87. These findings suggest a high level of interrater reliability due to the percent of variation across portfolios and for interactions within portfolios being close to zero (Stuhlmann, et al., 1999). Stuhlmann et al. (1999) further investigated whether implementing training on how to interpret the scoring rubric would further increase interrater reliability. The same rubric that was utilized in the pilot study was again employed in the second study.

The raters in the experimental group participated in an hour-long training session before school. The training included an explanation of each subscale in the rubric, specific examples that corresponded to each point criteria, practice on scoring written

examples, and compared scores for practiced items. Participants discussed comparison scores and researchers determined from the discussion which areas of the rubric were the most confusing. Further training was provided on the areas designated to be most confusing for teachers. The control group received no training and was asked to score student written work to the best of their ability utilizing the rubric scoring system.

The results indicated that there was no increase in interrater reliability between the experimental and control groups. However, further analysis did demonstrate that there was higher variability in scoring on four of the six subscales for the non-trained teachers. Therefore, the trained teachers were more consistent in their scoring on the subscales of pre-writing/picture, sentence, punctuation and capitalization, and format. The results did not suggest statistically significant differences for the subscales of story and mechanics/spelling between the control and experimental groups. Overall, Stuhlmann et al. (1999) concluded that the rubric training might have had a positive impact on the teachers' abilities to score student written samples more consistently within the rubric subscales.

### *Summary of Literature*

Overall, the literature indicated that theoretically, analytic rubrics are reliable and valid instruments for measuring the performance levels on a variety of assessment objects. However, there were contradictory and inconclusive results regarding whether holistic or analytic rubrics were the most effective scoring strategy. The pattern present in the research suggests that the type of performance level being assessed may be the best indicator of which rubric type would be most effective. However, the research did substantiate that the utilization of an analytic rather than a holistic rubric increased

interrater reliabilities within categories. Additionally, the research provides several experimental methodologies to evaluate rubrics for reliability, validity, and generalization.

### *Research Problem*

Although previous research provides a framework for rubric effectiveness and experimental methodologies to evaluate rubric reliability, the research has yet to specifically investigate objective assessment instruments for school psychology case studies. Consequently, due to an increased emphasis nationally on accountability in education, school psychology training programs will continue to require technically sound assessment tools to evaluate student performance on learning outcomes. The current study seeks to investigate the inter-rater agreement of an analytic rubric to assess performance on case studies conducted by school psychology trainees.



## Methods

### *Participants*

The participants in this study included 13 school psychology graduate students that were completing their 9-month internship. All of the school psychology interns had previously attended a 15-week course on conducting academic case studies and a 15-week course on conducting behavioral case studies. The participants consisted of nine female and 4 male students. Eleven of the participants were White, one participant was African American, and one participant was an international student from India. Seven participants had undergraduate degrees in psychology, five had undergraduate degrees in education, and one participant had a communication undergraduate degree. Two of the participants had previously earned a Ph.D. and eleven of the participants had previously earned a M.S. and were seeking the Education Specialist (Ed.S.) as their terminal degree.

### *Materials*

*Power Point Presentation.* A 39-slide power point presentation was the primary mode of training. The power point presentation described all five sections of the rubric and the overall rating section. The presentation provided key terms and definitions required to understand and apply the rubric criteria (Appendix B). The power point presentation was accompanied by a 32-page training manual. The training manual consisted of case study examples that corresponded to the five rubric sections and the overall rating section.

*Case Studies.* A total of nine case studies completed previously by school psychology students in a graduate program provided specific examples. The use of multiple case studies enabled the researcher to reduce the probability that the raters were biased by one particular case study, and to ensure that the rubric was consistently reliable

across case studies. Six case studies had an academic orientation and three case studies were behaviorally oriented. All case studies were modified from the original version. The academic and behavioral case studies were modified so that there were no academic and/or behavioral case studies uniquely represented the rubric criteria for 'outstanding', 'substantially developed', 'competent', 'threshold development', and 'needs development' (Appendix C).

*Case Study Evaluation Rubric.* The case study evaluation rubric was developed based upon best practices in school psychology. The rubric was developed over a 6-year period by a faculty member in the School Psychology Program. The rubric is composed of five sections and an overall rating section. The five sections consist of: local norms, problem identification and analysis, hypothesis testing, intervention, and evaluation and recommendations. The case study evaluation has five different rating levels: outstanding (all components in the competent and outstanding categories are checked), substantially developed (all components in the competent category plus some components in the outstanding category are checked), competent (all components in the competent category are checked), threshold development (some components in the competent category are checked), and needs development (one or more of the components in the needs development category is checked). Additionally, the case study is evaluated as an overall product based upon the following criteria: outstanding (case study is rated outstanding in all five sections), substantially developed (case study is rated competent or higher for all sections and substantially developed or higher in one or more sections), competent (all five sections of the case study are rated competent), threshold development (some but not

all sections are rated competent), and needs improvement (one or more of the sections is rated needs development) (See Appendix A).

Each domain category is comprised of an ordinal coding system that includes five scoring dimensions (choices). The case study rubric is categorized as an analytic rubric that is deemed by previous research to provide higher correlations and greater consistency among scorers, particularly within sub-areas.

Section one evaluates local norms. Within the local norms sections, the culture of the target classroom is analyzed and observed. This section addresses the class-wide rate of progress for the academic or behavioral targets and reviews the current curriculum. The classroom instructional methods are observed and a task analysis for the target behavior is conducted. Section two evaluates problem identification and analysis. Within this section, the at-risk student and the academic or behavioral concern is identified and specifically clarified. Additionally, a skill analysis and a performance analysis are conducted to determine if the student has a skill or performance deficit that is contributing to the target behavior. Hypothesis testing is evaluated in section three. Hypotheses are developed and tested to specifically identify the cause or source of the academic or behavioral target. In section four, intervention is evaluated based upon whether the intervention emerged logically from the previous sections. Intervention implementation and monitoring are also addressed within this section. Evaluation and recommendations comprise section five of the rubric. Data obtained from the intervention monitoring is utilized to determine the efficacy of the intervention. Furthermore, recommendations for additional educational or behavioral planning are discussed based upon the effectiveness of the intervention.

*Training of Raters*

The training was conducted on an intern seminar date for which all of the interns were scheduled to attend. The training was scheduled as part of the intern seminar agenda. The interns were given a description and purpose for the current study. The interns were given informed consent and an explanation that they were under no obligation to participate in the study. No incentives were given for participation.

Each participant initially engaged in the one-day, four hour training on rubric scoring. The training was divided into two sessions. The first half of the training, conducted over a 3-hour time period, consisted of a trainer providing definitions and examples for each domain on the rubric via a power point presentation. Additionally, the trainer discussed example case studies from the training manual with the trainees. The trainees engaged in scoring a sample case study with assistance from the trainer. This allowed for practice rubric scoring and discussion time in which the trainees could ask questions regarding scoring. The last half of the training conducted over a 2-hour time period and consisted of trainees scoring sample case studies.

*Research Design*

This study employed a correlational research approach to explore the relationship of the rubric ratings provided by 13 independent raters. Previous research investigating rubric inter-rater agreement conducted correlation coefficients. However, due to the ordinal nature of the data in this study, inter-rater agreement was calculated by dividing the number of agreed upon ratings by the total number of ratings for each of the nine case studies.

*Procedures*

Each month, participants were asked to evaluate additional case studies independently and submit them to the investigator. Participants anonymously submitted case studies to the investigator to ensure confidentiality. The number of case studies completed significantly varied over time. All 13 participants completed case studies during the initial training. However, in the subsequent months the return rate decreased to 19%. Seventy-two of the possible 117 case studies were submitted to the investigator. The return rate was 62% for the nine case studies that were evaluated.

*Data Analysis*

Given the ordinal nature of the data, independent ratings were cross-tabulated for each section of the Case Study Rubric and the overall rating. The percentage of inter-rater agreement was calculated by dividing the agreed upon ratings by the total number of ratings across the nine case studies for each section and for the overall rating.

## Results

Data analyses were conducted on a total of nine case studies, 6 academic case studies and 3 behavioral case studies. Individual ratings for each of the five sections were reviewed to determine inter-rater agreement. The ratings for the overall case study were reviewed in the same manner. Each rating was identified to be accurate or inaccurate and a frequency count was completed for each category. A descriptive summary of the data was supplemented with a determination of the percent of inter-rater agreement by using the formula:  $N \text{ of accurate ratings} / (N \text{ of accurate ratings} + N \text{ of inaccurate ratings}) \times 100$ .

The results included 332 agreed upon ratings. Therefore, the rubric was demonstrated to have a 66% inter-rater agreement for the 5 sections and the overall rating. Levels of agreement across ratings for each of the five sections ranged from 52% to 78%. Section 1: Local Norms had the lowest rate of agreement with 52%, while Section 4: Intervention had the highest rate of agreement with 78%. Section 2: Problem Identification had a 62% rate of agreement, Section 3: Hypothesis Testing rate of agreement was 68%, and Section 5: Evaluation had a 70% rate of agreement. Overall ratings of the case studies had an inter-rater agreement of 74%.

Table 1

### Percentage of Inter-rater Agreement

Section	Inter-rater Agreement
Local Norms	52%
Problem Identification	62%
Hypothesis Testing	68%

Section	Inter-rater Agreement
Intervention	78%
Evaluation	70%
Overall Rating	74%
Five Sections and Overall Rating	66%

## Discussion

The results indicate that sections 1 and 2 will need to be reviewed and adjusted to increase the inter-rater agreement of this rubric. The results of sections 3, 4, and 5, and the overall rating indicate that a higher level of confidence can be assumed with regard to an evaluator's ability to detect compliance with these sections. Further studies would be beneficial to evaluate the rubric inter-rater agreement once sections 1 and 2 are reviewed and adjusted.

Additional studies could include participants with less training in conducting student case studies. Further research would assist in determining whether the inter-rater agreement is generalizable across the school psychology spectrum of trainers, trainees, and practitioners.

The case study rubric can be utilized by a variety of professionals in the fields of school psychology and special education. The rubric can provide an outline for trainees to conduct student case studies and is an invaluable assessment tool for school psychology trainers. The rubric provides trainers and training programs with an evaluation tool to objectively assess student performance on conducting case studies. Furthermore, the rubric is designed so that case studies that meet the criteria as designated on the rubric could exemplify demonstration of knowledge skills in each of the 11 National Association of School Psychologists Standard Domains (except domain 2.10).

The case study rubric could also lay the foundation for introducing the concept of 'response to intervention' (RTI) to school districts and special education personnel. The rubric provides an outline for assessing local norms, skill/behavior deficits, intervention



implementation, and data collection. Therefore, the rubric could be utilized as an outline for conducting case studies in level three of the RTI process.

Overall, the case study rubric can be utilized by numerous professionals in the fields of school psychology and special education. Further research would be beneficial to validate the inter-rater agreement for the various usages other than assessing student performance on conducting student case studies. The results of the current evaluation suggest that this rubric can provide a structured, objective measure of performance by school psychology students on case studies.

## References

- Abedi, J. & Baker, E.L. (1995). A latent-variable modeling approach to assessing interrater reliability, topic generalizability, and validity of a content assessment scoring rubric. *Educational and Psychological Measurement*, 55(5), 701-715.
- Anderson, R.S. & Puckett, J.B. (2003). Assessing students' problem-solving assignments. *New Directions For Teaching and Learning*, 95, 81-87.
- Baker, E.L., Abedi, J., Linn, R.L., & Niemi, D. (1995). Dimensionality and generalizability of domain-independent performance assessments. *The Journal of Educational Research*, 89(4), 197-205.
- Boulet, J.R., Rebbecca, T.A., Denton, E.C., McKinley, D.W., & Whelan, G.P. (2004). Assessing the written communication skills of medical school graduates. *Advances in Health Science Education*, 9, 47-60.
- Choinski, E., Mark, A.E. & Murphey, M. (2003). Assessment with rubrics: An efficient and objective means of assessing student outcomes in an information resource class. *Libraries and the Academy*, 3(4), 563-575.
- Crehan, K.D. & Hudson, R. (2001). A comparison of two scoring strategies for performance assessments. *Educational Research Quarterly*, 25(2), 52- 55.
- Johnson, R.L., Fisher, S., Willeke, M.J., & McDaniel, F. (2003). Portfolio assessment in a collaborative program evaluation: The reliability and validity of a family literacy portfolio. *Evaluation and Program Planning*, 26, 367-377.
- Klein, S.P., Stecher, B.M., Sahvelson, R.J., McCaffrey, D., Ormseth, T., Bell, R.M., Comfort, K., & Othman, A.R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurements in Education*, 11(2), 121-137.
- Mabry, L. (1999). Writing to the Rubric: Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappan*, May, 673-679.
- National Association of School Psychologists. (2000). *Standards for Training and Field Placement Programs in School Psychology*. Retrieved October 2, 2004, <http://www.nasponline.org/certification/finalstandards.html>.
- Novak, J.R., Herman, J.L., & Gearhart, M. (1996) Establishing validity for performance based assessments: An illustration for collections of student writing. *The Journal of Educational Research*, 89(4), 220-233.
- Penny, J., Johnson, R.L., & Gordon, B. (2000a). Using rating augmentation to expand

the scale of an analytic rubric. *Journal of Experimental Education*, 68(3), Spring, 269-288.

Penny, J., Johnson, R.L., & Gordon, B. (2000b). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing*, 7, 143-164.

Stuhlmann, J., Daniel, C., Dellinger, A., Denny, R.K., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Journal of Reading Psychology*, 20, 107-127.

## Appendix A

## Case Study Rubric

## Section 1.0

**Local Norms:** Local norms and outcome goals were established for class.

	Outstanding	Competent	Needs Development
1.1	<input type="checkbox"/> Teacher consultation provided classwide behavioral and/or academic goals and a target date to accomplish the classwide goals		
1.2	<input type="checkbox"/> The class goal statement(s) was written in observable, measurable terms, and was based on the all of the following: <ul style="list-style-type: none"> <li><input type="checkbox"/> Review of curriculum for academic goals, AND</li> <li><input type="checkbox"/> Task analysis for academic and/or behavioral target goals, AND</li> <li><input type="checkbox"/> Description of class-wide instructional methods to address the academic and/or behavioral target goals</li> </ul>	<input type="checkbox"/> The class goal statement(s) was written in observable, measurable terms	<input type="checkbox"/> The class goal statement(s) was NOT written in observable, measurable terms
1.3	<input type="checkbox"/> Local norms were established through direct observation, criteria-based instrument(s), or curriculum-based measurement (Classes that do not have established local norms will need to have at least 3 administrations of each measure conducted over a several week period to determine average rate of change per week or stability for class.)	<input type="checkbox"/> Local norms were established through direct observation, criteria-based instrument(s), or curriculum-based measurement	<input type="checkbox"/> Local norms and/or goals were underdeveloped
1.4	<input type="checkbox"/> Technology was used in the gathering and synthesis of data		

## Rating for 1.0

<input type="checkbox"/> <b>Outstanding:</b> All components in the Competent and Outstanding categories are checked	<input type="checkbox"/> <b>Substantially Developed:</b> All components in the Competent category plus some components in the Outstanding category are checked	<input type="checkbox"/> <b>Competent:</b> All components in the competent category are checked	<input type="checkbox"/> <b>Threshold Development:</b> Some components in the competent category are checked	<input type="checkbox"/> <b>Needs Development:</b> One or more of the components in the Needs Development category is checked
--	--	---	--	---

**Section 2.** **Problem Identification & Analysis:** The at-risk student and academic/behavioral concern(s) are identified and clarified.

	<b>Outstanding</b>	<b>Competent</b>	<b>Needs Development</b>
2.1		<input type="checkbox"/> One at-risk student is identified	
2.2		<input type="checkbox"/> The at-risk student's academic and/or behavioral concern(s) is identified and operationally defined using class goals and local norms	<input type="checkbox"/> The at-risk student's academic and/or behavioral concern(s) is identified but NOT operationally defined using class goals and local norms
2.3		<input type="checkbox"/> The problem is identified and defined collaboratively	
2.4		<input type="checkbox"/> A baseline for the at-risk student is established for the concern(s)	<input type="checkbox"/> A baseline for the at-risk student is NOT established or is inappropriate
2.5	<input type="checkbox"/> <u><b>Skill analysis</b></u> was conducted and included <u><b>all of the components listed under Competent</b></u>	<input type="checkbox"/> <u><b>Skill analysis</b></u> was conducted and included <u><b>one or more</b></u> of the following: <input type="checkbox"/> Error analysis, <input type="checkbox"/> Direct observation of skill, <input type="checkbox"/> Criteria-based assessment, OR curriculum-based assessment	<input type="checkbox"/> No skill analysis was conducted, or analysis was inappropriate for the identified concern(s)
2.6	<input type="checkbox"/> <u><b>Performance analysis</b></u> was conducted and included <u><b>all of the components listed under Competent</b></u>	<input type="checkbox"/> <u><b>Performance analysis</b></u> was conducted and included <u><b>one or more</b></u> of the following: <input type="checkbox"/> Record review for historical documentation of pertinent information, <input type="checkbox"/> Student interview, <input type="checkbox"/> Ecological or situational analysis of concern (e.g., routines, expectation-skill match, relationships, classroom environment, adult/teacher support, cultural issues) <input type="checkbox"/> Direct observation (e.g., on-task) <input type="checkbox"/> Parent interview	<input type="checkbox"/> No performance analysis was conducted, or analysis was inappropriate for the identified concern(s)

<b>Rating for 2.0</b>				
<input type="checkbox"/> <u><b>Outstanding:</b></u> All components in the Competent and Outstanding categories are checked	<input type="checkbox"/> <u><b>Substantially Developed:</b></u> All components in the Competent category plus some components in the Outstanding category are checked	<input type="checkbox"/> <u><b>Competent:</b></u> All components in the competent category are checked	<input type="checkbox"/> <u><b>Threshold Development:</b></u> Some components in the competent category are checked	<input type="checkbox"/> <u><b>Needs Development:</b></u> One or more of the components in the Needs Development category is checked

## Section 3.0

**Hypothesis Testing:** Hypotheses were developed and tested

	Outstanding	Competent	Needs Development
3.1	<input type="checkbox"/> Hypotheses were generated through collaboration with teacher and/or parent		
3.2	<input type="checkbox"/> Multiple hypotheses were developed to identify the cause or source of each problem	<input type="checkbox"/> One hypothesis was developed to identify the cause or source of each problem	<input type="checkbox"/> No hypotheses were developed
3.3	<input type="checkbox"/> Each hypothesis was tested to confirm the cause or source of the problem using one or more of the following methods: <input type="checkbox"/> Direct observation, <input type="checkbox"/> Analogue assessment, <input type="checkbox"/> Functional assessment, <input type="checkbox"/> Self-monitoring assessment, <input type="checkbox"/> Other	<input type="checkbox"/> The hypothesis was tested to confirm the cause or source of the problem using one or more of the following methods: <input type="checkbox"/> Direct observation, <input type="checkbox"/> Analogue assessment, <input type="checkbox"/> Functional assessment, <input type="checkbox"/> Self-monitoring assessment, <input type="checkbox"/> Other	<input type="checkbox"/> Hypothesis testing did not occur
3.4		<input type="checkbox"/> The hypothesis reflected awareness of individual differences (e.g., biological, social, linguistic, cultural)	
3.5		<input type="checkbox"/> Hypothesis testing linked the academic and/or behavioral problem(s) with the intervention	<input type="checkbox"/> Hypothesis testing did NOT link the academic and/or behavioral problem(s) with the intervention

Rating for 3.0				
<input type="checkbox"/> <b>Outstanding:</b> All components in the Competent and Outstanding categories are checked	<input type="checkbox"/> <b>Substantially Developed:</b> All components in the Competent category plus some components in the Outstanding category are checked	<input type="checkbox"/> <b>Competent:</b> All components in the competent category are checked	<input type="checkbox"/> <b>Threshold Development:</b> Some components in the competent category are checked	<input type="checkbox"/> <b>Needs Development:</b> One or more of the components in the Needs Development category is checked

**Section 4. Intervention:** Intervention was implemented and monitored

	Outstanding	Competent	Needs Development
4.1		<input type="checkbox"/> Goal statement(s) was written in observable, measurable terms	<input type="checkbox"/> Goal statement was NOT written in observable, measurable terms
4.2		<input type="checkbox"/> Goal statement(s) emerged from the problem analyses and hypothesis testing	
4.3		<input type="checkbox"/> Intervention(s) was developed collaboratively	<input type="checkbox"/> Intervention(s) was NOT developed collaboratively
4.4		<input type="checkbox"/> Intervention(s) logically linked to the referral question	<input type="checkbox"/> Intervention was NOT linked to referral question
4.5		<input type="checkbox"/> Intervention(s) logically linked to the hypothesis	
4.6		<input type="checkbox"/> Intervention(s) logically linked to the goal statement	
4.7	<input type="checkbox"/> Intervention(s) was described including procedures for one or more of the following: <input type="checkbox"/> Promoting new or replacement behaviors/skills <input type="checkbox"/> Increasing existing behaviors/skills <input type="checkbox"/> Reducing interfering problem behaviors <input type="checkbox"/> Facilitating generalization	<input type="checkbox"/> Intervention(s) was described in enough detail to ensure appropriate implementation	
4.8	<input type="checkbox"/> Support was provided to justify the use of the intervention as evidence-based practice (e.g., research literature, functional analysis)		<input type="checkbox"/> Intervention(s) was limited to determination of eligibility for special education services
4.9	<input type="checkbox"/> Acceptability of intervention by teacher, parent and child was verified	<input type="checkbox"/> Intervention reflected sensitivity to individual differences, resources, classroom practices, and other system issues	
4.10	<input type="checkbox"/> Logistics of setting, time, resources and personnel required for intervention and data gathering were defined and implemented	<input type="checkbox"/> Intervention(s) was implemented	<input type="checkbox"/> Intervention(s) was limited to referral for services external to the school and/or the home
4.11	<input type="checkbox"/> Treatment/intervention integrity was monitored to assure appropriate implementation	<input type="checkbox"/> Intervention(s) was monitored	

Rating for 4.0				
<input type="checkbox"/> <b><u>Outstanding:</u></b> All components in the Competent and Outstanding categories are checked	<input type="checkbox"/> <b><u>Substantially Developed:</u></b> All components in the Competent category plus some components in the Outstanding category are checked	<input type="checkbox"/> <b><u>Competent:</u></b> All components in the competent category are checked	<input type="checkbox"/> <b><u>Threshold Development:</u></b> Some components in the competent category are checked	<input type="checkbox"/> <b><u>Needs Development:</u></b> One or more of the components in the Needs Development category is checked



## Section 5.0

**Evaluation and Recommendations:** Data were gathered and documented to demonstrate efficacy of intervention.

	Outstanding	Competent	Needs Development
5.1	<input type="checkbox"/> Goal attainment was plotted at the end point and compared to baseline	<input type="checkbox"/> Progress monitoring data were plotted on a graph or chart	<input type="checkbox"/> Progress monitoring data were NOT plotted on a graph or chart
5.2	<input type="checkbox"/> Goal attainment was plotted at the end point and compared to the desired goal	<input type="checkbox"/> Data were provided as evidence of measurable, positive impact toward stated goal	<input type="checkbox"/> Data were NOT provided to document student progress
5.3	<input type="checkbox"/> Single-case design was specified (e.g., changing criterion, withdrawal, multiple baseline, alternating treatments) to prove efficacy of intervention		
5.4	<input type="checkbox"/> Current technologies were used to present data		
5.5	<input type="checkbox"/> Data were obtained through multiple methods and were presented in support of student's progress from two or more of the following: <input type="checkbox"/> Direct observation <input type="checkbox"/> Rating scale <input type="checkbox"/> Peer comparison <input type="checkbox"/> Self-monitoring <input type="checkbox"/> CBM <input type="checkbox"/> Other	<input type="checkbox"/> Evidence in support of student's progress from one of the following: <input type="checkbox"/> Direct observation <input type="checkbox"/> Rating scale <input type="checkbox"/> Peer comparison <input type="checkbox"/> Self-monitoring <input type="checkbox"/> CBM <input type="checkbox"/> Other	
5.6	<input type="checkbox"/> Intervention quality and integrity were monitored with a formal measure		
5.7	<input type="checkbox"/> Effectiveness of intervention was examined collaboratively		<input type="checkbox"/> Effectiveness of intervention was NOT discussed
5.8	<input type="checkbox"/> Intervention limitations and side effects were described		
5.9	<input type="checkbox"/> Strategies for follow-up were developed collaboratively	<input type="checkbox"/> Suggestions for follow-up were provided	

## Rating for 5.0

<input type="checkbox"/> <b><u>Outstanding:</u></b> All components in the Competent and Outstanding categories are checked	<input type="checkbox"/> <b><u>Substantially Developed:</u></b> All components in the Competent category plus some components in the Outstanding category are checked	<input type="checkbox"/> <b><u>Competent:</u></b> All components in the competent category are checked	<input type="checkbox"/> <b><u>Threshold Development:</u></b> Some components in the competent category are checked	<input type="checkbox"/> <b><u>Needs Development:</u></b> One or more of the components in the Needs Development category is checked
---	---	--	---	--

Overall Rating for Case Study (A rating of Competent or higher is required to pass)				
<input type="checkbox"/> <b>Outstanding:</b> Case study is rated Outstanding in all five Sections	<input type="checkbox"/> <b>Substantially Developed:</b> Case study is rated Competent or higher for all Sections and Substantially Developed or higher in one or more sections	<input type="checkbox"/> <b>Competent:</b> All five Sections of the Case Study are rated competent	<input type="checkbox"/> <b>Threshold Development:</b> Some but not all Sections are rated Competent	<input type="checkbox"/> <b>Needs Development:</b> One or more of the Sections is rated Needs Development

*Appendix B*

Training: Case Study Evaluation Rubric  
 Case Study – 5 Stage Problem Solving Model

■ Stage 1 – Local Norms

- The culture of the target classroom is analyzed and observed.
- Class-wide rate of progress for academic or behavioral targets is determined
- Review of curriculum is conducted
- Classroom instructional methods are observed
- Task analysis for academic or behavioral targets is conducted

Case Study – 5 Stage Problem Solving Model

■ Stage 2 – Problem Identification & Analysis

- The at-risk student and the academic or behavioral concern(s) are identified and specifically clarified
- Skill analysis is conducted
- Performance analysis is conducted

Case Study – 5 Stage Problem Solving Model

■ Stage 3 – Hypothesis Testing

- Hypotheses are developed and tested to identify the cause or source of the academic or behavioral target

Case Study – 5 Stage Problem Solving Model

■ Stage 4 – Intervention

- The intervention is implemented and monitored
- The intervention to be implemented emerged from results obtained during hypothesis testing

Case Study – 5 Stage Problem Solving Model

■ Stage 5 – Evaluation & Recommendations

- Data obtained from intervention monitoring is utilized to determine the efficacy of the intervention
- Recommendations for further educational or behavioral planning are discussed based upon the efficacy of the intervention

Case Study Evaluation - Rubric

■ The rubric has 5 problem solving stages that were just described: local norms, problem identification & analysis, hypothesis testing, intervention, and evaluation & recommendations

■ Each stage is assessed by listed criteria that correspond to five different ratings:

- Outstanding
- Substantially Developed (has all characteristics for Competent Development and some characteristics for Outstanding rating)
- Competent Development
- Threshold Development (has some characteristics for Competent rating)
- Needs Development

### Outstanding – Stage 1: Local Norms

■1.1 Teacher consultation provided class-wide behavioral and/or academic goals and a target date to accomplish the class-wide goals

–Definition – the teacher reports an academic or behavioral goal for the classroom and a date in which the teacher would like to accomplish the goal.

### Outstanding & Competent Development – Stage 1: Local Norms

■1.2 The class goal statement was written in observable, measurable terms

–Definition – the class goal statement was devised so that data could be collected to determine the rate of progress for the goal

### Outstanding & Competent Development – Stage 1: Local Norms

■1.3 Local norms were established through direct observation, criteria-based instrument, or curriculum-based measurement

–Definition – The class-wide rate of progress for either academic or behavioral targets are determined.

■(Classes that do not have local norms already established will need to have at least 3 measures conducted over a several week period to determine average rate of change per week or stability for class) / criteria for outstanding

### Outstanding – Stage 1: Local Norms

■1.4 Technology was used in the gathering and synthesis of data

–Definition –

Data is presented in a graph.

### Outstanding & Competent Development – Stage 2: Problem Identification & Analysis

■2.1 At-risk student was identified

–Definition – the at-risk student to receive intervention was selected

### Outstanding & Competent Development – Stage 2: Problem Identification & Analysis

■2.2 At-risk student's academic and/or behavioral concern(s) was identified and operationally defined using class goals and local norms

–Definition – The at-risk student's targeted behavior is defined in measurable terms and a goal is set based upon class norms.

### Outstanding & Competent Development – Stage 2: Problem Identification & Analysis

■2.3 Problem was identified and defined collaboratively

–Definition – The teacher and/or parent were involved in identifying and defining the problem.

### Outstanding & Competent Development – Stage 2: Problem Identification & Analysis

■2.4 Baseline for at-risk student was established for concern(s)

–Definition – The student is assessed to determine his/her current rate of performance for the academic or behavioral target.

### Outstanding & Competent Development – Stage 2: Problem Identification & Analysis

■2.5 Skill analysis

–Definition – the identified student's target behavior is evaluated to determine the source of the problem – is there a skill deficit?

–The skill analysis consists of one or more of the following: a. Error analysis b. Direct observation of skill c. Criteria-based assessment OR curriculum-based assessment (must include all components to meet outstanding criteria)

Outstanding & Competent Development – Stage 2: Problem Identification & Analysis  
2.6 Performance analysis

–Definition - the identified student's target behavior is evaluated to determine the source of the problem – is there a performance issue?

–One or more of the following is conducted:

- a. Record review of historical documentation of pertinent information
- b. Student interview
- c. Ecological or situational analysis of concern
- Direct observation
- Parent interview
- (all components must be included to meet outstanding criteria)

Outstanding & Competent Development – Stage 3: Hypothesis Testing

3.1 Hypotheses were generated through collaboration with teacher and/or parent (outstanding criteria)

3.2 One hypothesis were developed to identify the cause or source of the problem

–Definition –1 hypothesis was developed to determine source of academic or behavioral targets

–More than 1 hypothesis is required for outstanding criteria

Outstanding & Competent Development – Stage 3: Hypothesis Testing

■3.3 The hypothesis was tested to confirm the cause or source of each problem using one or more of the following methods:

- Direct observation
- Analogue assessment
- Functional analysis
- Self-monitoring assessment
- Other

Outstanding & Competent Development – Stage 3: Hypothesis Testing

■3.4 Hypotheses reflected awareness of individual differences (biological, social, linguistic, cultural)

–Definition – The hypotheses take into consideration individual differences among students.

Outstanding & Competent Development – Stage 3: Hypothesis Testing

■3.5 Hypothesis testing linked the academic and/or behavioral problem(s) with the intervention

–Definition – Results obtained from hypothesis testing provided information about what type of intervention should be implemented

Outstanding & Competent Development – Stage 4: Intervention

■4.1 Goal statement was written in observable, measurable terms

–Definition – The student's goal for the target behavior was written so that data collected could determine rate of progress

Outstanding & Competent Development – Stage 4: Intervention

■4.2 Goal statement(s) emerged from the problem analysis and hypothesis testing

–Definition – The student's academic or behavioral goal was based upon results obtained during the problem analysis phase and hypothesis testing phase

Outstanding & Competent Development – Stage 4: Intervention

■4.3 Intervention was developed collaboratively

–Definition – Intervention was developed with the student's teacher, student's parent, and student.

Outstanding & Competent Development – Stage 4: Intervention

■4.4 Intervention logically linked to the referral question

–Definition – the intervention to be implemented was directly related to initial concerns regarding target student's academic or behavioral concerns

Outstanding & Competent Development – Stage 4: Intervention

■4.5 Intervention logically linked to the hypothesis

–Definition – The intervention to be implemented was based upon hypothesis testing results

Outstanding & Competent Development – Stage 4: Intervention

■4.6 Intervention logically linked to the goal statement

–Definition – the intervention to be implemented is directly related to the goal set for academic or behavioral target

Outstanding & Competent Development – Stage 4: Intervention

■4.7 (COMPETENT) Intervention was described in enough detail to ensure appropriate implementation

–Definition – The intervention was described so that the person implementing the intervention had a clear understanding of how to conduct the intervention.

■ (OUTSTANDING) Intervention was described including procedures for one or more of the following:

–Promoting new/replacement behavior/skills

–Increasing existing behavior/skills

–Reducing interfering problem behaviors

–Facilitating generalization

Outstanding – Stage 4: Intervention

■4.8 Support was provided to justify the use of the intervention as evidence-based practice (e.g. research literature, functional analysis)

–Definition – Research based support was provided for intervention efficacy.

Outstanding & Competent Development – Stage 4: Intervention

■4.9 (COMPETENT) Intervention(s) reflected sensitivity to individual differences, resources, classroom practices, and other system issues

■ (OUTSTANDING) Acceptability of intervention by teacher, parent, and child was verified (outstanding criteria)

–Definition – the intervention plan designed taking into account the culture of the student's classroom and school

Outstanding & Competent Development – Stage 4: Intervention

■4.10 (COMPETENT) Intervention was implemented

■(OUTSTANDING) Logistics of setting, time, resources and personnel required for intervention and data gathering were defined and implemented

Outstanding & Competent Development – Stage 4: Intervention

■4.11 (COMPETENT) Intervention was monitored

■(OUTSTANDING) Treatment integrity was monitored to assure appropriate implementation

Outstanding and Competent Development – Stage 5: Evaluation and Recommendations

5.1 (COMPETENT) Progress monitoring data and goal attainment data plotted on a graph or chart

(OUTSTANDING) Goal attainment was plotted at end point and compared to baseline

–Definition – Intervention results are presented in a visual graphic format

–Definition – Progress from intervention results are compared to baseline

Outstanding & Competent Development – Stage 5: Evaluation & Recommendations

■ 5.2 (COMPETENT) Data were provided as evidence of measurable, positive impact toward stated goal (criteria for competent development)

■(OUTSTANDING) Goal attainment was plotted at end point and compared to the desired goal

–Definition – Progress from intervention results are compared to the target goal in a visual graphic representation

Outstanding – Stage 5: Evaluation & Recommendations

■5.3 Single-case design was specified (e.g. changing criterion, withdrawal, multiple baseline, alternating treatments) to prove efficacy of intervention

■5.4 Current technologies were used to present data

Outstanding & Competent Development – Stage 5: Evaluation & Recommendations

Outstanding – Stage 5: Evaluation & Recommendations

5.6 Intervention quality and integrity were monitored with a formal measure

5.7 Effectiveness of intervention was examined collaboratively

–Definition – the effectiveness of the intervention is discussed with the teacher, parent, and student

5.8 Intervention limitations and side effects were described

Competent Development – Stage 5: Evaluation & Recommendations

■5.9 (COMPETENT) Suggestions for follow-up were provided

–Definition – Further interventions to continue the student's current rate of progress were provided or adaptations to the intervention were recommended.

■(OUTSTANDING) Strategies for follow-up were developed collaboratively

–Definition – follow-up plans and recommendations are developed by the teacher, school psychology assistant, and parent

■Questions?

## Appendix C

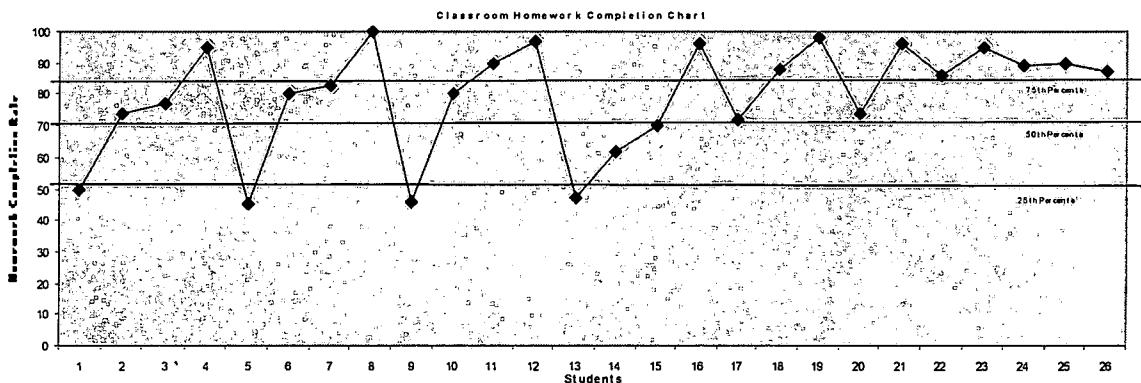
### Problem Identification

The science and math teacher indicated that her class wide behavioral goal was for her students to complete their homework on a weekly basis 90% of the time by February 28<sup>th</sup>. The teacher provided information regarding the district's policies for behavior. The district had adopted a new behavior program in which teachers monitored homework completion, office referrals, and detentions. Students who turned in their homework 90% of the time, has 0% of office referrals and detentions was rewarded on a weekly basis. Therefore, due to the new district program, the teacher monitored the percent of homework completion, office referrals, and detentions for each student. The students were made aware of their percentages in all three areas on a weekly basis. Students' grades in conduct were derived from their rate of homework completion, office referrals, and detentions.

Teacher interview indicated that she conferences daily and weekly with each student regarding their rate of homework completion, office referrals, and detentions. Therefore, students are reminded daily of their current standing and eligibility for reward at the end of the week.

During the interview, the classroom teacher reviewed her grade book and calculated that her students were turning in 85% of their homework in a weeks time. Four students were performing below the 25<sup>th</sup> percentile, 5 students were performing between the 25<sup>th</sup> and 50<sup>th</sup> percentiles, 10 students were performing between the 50<sup>th</sup> and 75<sup>th</sup> percentiles and 7 students were performing above the 75<sup>th</sup> percentile.

Review of the teacher grade book also indicated that the students were completing 85% of their work per week during the past month. The homework completion rate appeared to be fairly stable over the past 4 weeks. Four students were performing below the 25<sup>th</sup> percentile, 5 students were performing between the 25<sup>th</sup> and 50<sup>th</sup> percentiles, 10 students were performing between the 50<sup>th</sup> and 75<sup>th</sup> percentiles and 7 students were performing above the 75<sup>th</sup> percentile. An analysis of documentation for students conduct grades also supported the above findings. The graph below provides a visual representation of the class percentiles.





### Problem Analysis

An analysis of the problem situation incorporated the use of the following assessment measures: the problem-solving interview (Barnett & Carey, 1992), assessment checklist for math problems, Behavioral Observation of Students in Schools (B.O.S.S.) (Shapiro, 1996), TIES Student Interview (Ysseldyke & Christenson, 1987), and curriculum-based measures of math computation (Howell, Fox, & Morehead, 1993; Shapiro, 1996).

Baseline measures obtained from teacher interview and record review indicated that Carlos was completing 50% of his homework per week, therefore he is performing at the 25<sup>th</sup> percentile. Carlos' teacher indicated that she would like for him to complete at least 90% of his work on a weekly basis.

Teacher observations suggested that Carlos had the ability to do grade level work, yet often did not complete and turn in his homework. The teacher stated that occasionally Carlos would complete and turn in work and when he did the work would receive grades of "A". The teacher also indicated that when Carlos did earn grades in Science, Math, and Language Arts, he would earn "A's". The teacher believed that Carlos did not have difficulty with the level of the academic content.

Classroom observations of Carlos' academic performance on work completed during class time demonstrated that he had the ability to work independently with few mistakes. Carlos would occasionally turn in homework which suggested that he did comprehend the process of turning in his homework. Lastly, a review of Carlos' test grades also verified that he did not have difficulty with the level of the work. Carlos received grades of "A" or "B" on all of his tests.

Carlos is a sixth grade student who has attended the same elementary school since kindergarten. He lives with his mother, father, and new baby sister. A review of his cumulative file indicates no medical concerns. Carlos passed his hearing and vision screenings during 5<sup>th</sup> grade. Further record review indicates that Carlos has a history of making "A's" and "B's" from kindergarten through 6<sup>th</sup> grade. A review of the teacher's grade book did not establish a pattern for when Carlos was most likely to turn in his homework.

During the teacher interview, Carlos' teacher indicated that at the beginning of the year, Carlos turned in the majority of his homework. The teacher did indicate that Carlos did interact with his peers and appeared to get along well with others in the classroom. His teacher did see a change in the rate of homework completion from the beginning of the year; however, no noted changes were indicated.

Interview with Carlos' mother indicated that she was very surprised to hear that Carlos was not turning in his homework. She reported that Carlos has never had a history or difficulty with school in the past. His mother stated that Carlos had always been a good student. The interview suggested that Carlos' mother just had a new baby 3 months old and wondered if that was affecting Carlos not turning in his homework.

Three Behavioral Observation of Students in Schools (BOSS) was used to record direct observations of Carlos during science, reading, and math instruction. During science, Carlos was observed to be actively engaged 60% of the time and passively engaged 30% of the time. The peer group comparison

was actively engaged 67% of the time and passively engaged 23% of the time. Carlos demonstrated off-task motor behaviors 15% of the time, off-task verbal behaviors 5% of the time, and off-task passive behaviors 0% of the time. Peer comparison off-task behaviors were consistent with Carlos' behavior.

In reading, Carlos was actively engaged 70% of the time, passively engaged 30% of the time, with on off-task behaviors noted. The peer comparison group's engagement time was again consistent with Carlos' time of engagement. Lastly, in math, Carlos was again actively engaged 70% of the time, passively engaged 30% of the time, with only minimal off-task behaviors. His peer comparison was similar.

The first hypothesis was developed with the teacher and parent and suggested that Carlos was not completing his homework because he did not want to ask his mother for help in the evenings since she needed to take care of his new baby sister. To test the hypothesis, Carlos' mother agreed to set aside 1 hour every night to work with Carlos on his homework. Hypothesis testing indicated that Carlos' mother fulfilled her obligations and documented that Carlos completed all of his homework on a nightly basis. However, teacher interview indicated that Carlos was still only turning in 50% of his work on a weekly basis.

The second hypothesis generated with collaboratively with teacher and parent was that Carlos did not have sufficient reinforcement for homework completion. The target behavior was Carlos's homework completion. The goal of 100% work completion was set by Carlos. In order to test the hypothesis, a behavioral contract was developed with Carlos. If Carlos completed his homework 100% of the time over a two-week period, he could have pizza for lunch from a local pizza restaurant. Hypothesis testing indicated that Carlos completed his work 100% of the time for the two-week period and was rewarded with pizza for lunch. Due to the significant results of the hypothesis testing, the above hypothesis was accepted. Additionally, the behavior contracting utilized for hypothesis testing provided the basis for the behavioral intervention.

A record review and an interview with Carlos' mother indicated that there were not medical or physical concerns. His primary language was English and his social and cultural experiences were similar to those of his peer group in his town.

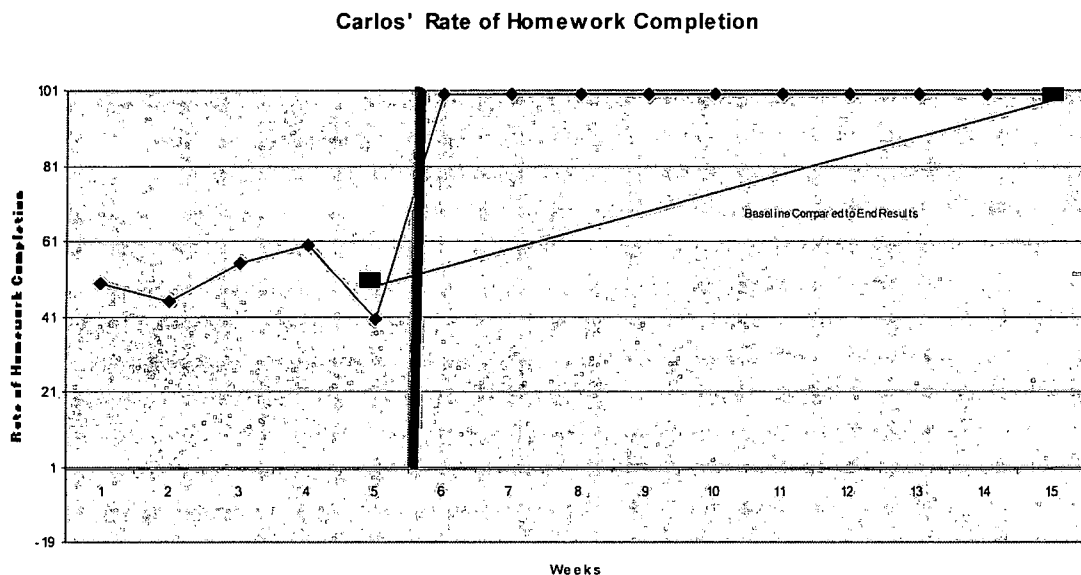
### Plan Implementation

I was the primary change agent in this intervention. A behavioral contract was developed with Carlos to increase reinforcing contingencies for homework completion. According to the contract, Carlos was required to complete 100% of his weekly homework. Self-selected reinforcement (pizza) was provided contingent upon three weeks of successful homework completion. The selected intervention was based upon Shapiro, et al, (1999) which indicated that behavioral contracting and reinforcing contingencies are empirically based and effective interventions to utilize for motivation issues. Monitoring of weekly homework completion was based on the teacher's verbal report and daily charting the following Monday. Treatment integrity was also monitored through the daily and weekly charts that reported Carlos' percent of homework completion. In addition, when Carlos earned reward, a note was added to the

weekly chart indicating if he received the pizza reward. The contract was in effect for eight weeks.

### Plan Evaluation

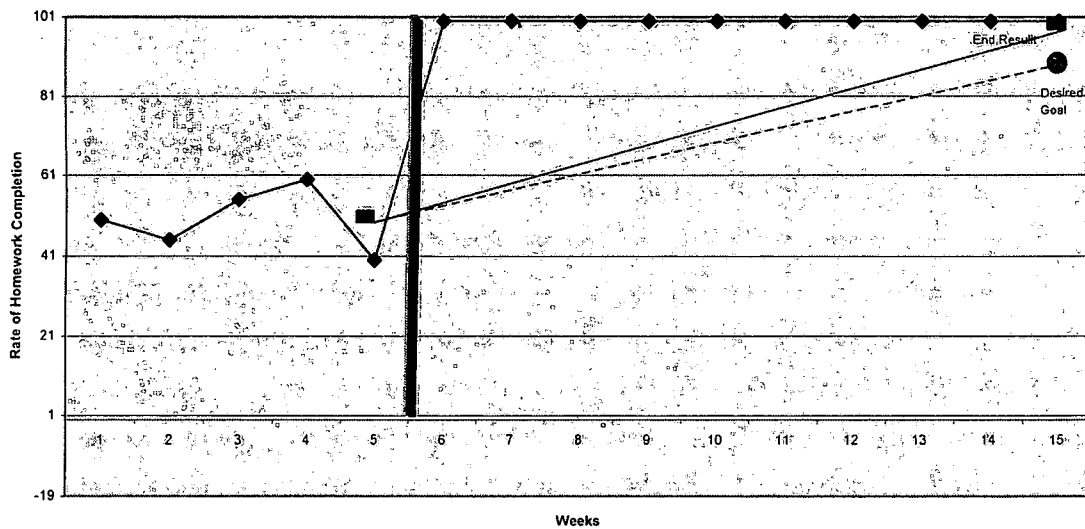
The effectiveness of the intervention to increase homework completion was determined using a single-case, time series analysis design. The percentage of homework assignment completed weekly was graphed over time. After eight school weeks of implementing the behavioral contract, Carlos's homework completion rate was found to be at a significantly higher level of 99% relative to his baseline performance of 50%. At this point the intervention was judged to be successful and fading was introduced. The first step in fading included the removal of the self-selected reinforcement for homework completion. Below is a graph that provides visual information regarding Carlos' progress during the intervention. The graph compares his baseline point to the actual end result of the intervention.



Carlos's teacher was asked to complete an intervention rating profile to measure the degree to which she found the goals, procedures, and outcomes of the behavior intervention acceptable (social validity). On a 6-point Likert scale where 1 equals "Strongly Disagree" and 6 equals "Strongly Agree," Carlos's teacher reported ratings of 5 or better to items regarding the degree to which she judged the intervention was beneficial for Carlos's, the intervention to be appropriate for a variety of children, and if teachers would be likely to use the because it required little training to implement effectively. The following three items received ratings of 2 or lower: this intervention being disruptive to other children, this intervention would have negative effects on other children, this intervention would be difficult to implement in a classroom with 30 other students.

A review of the teacher's weekly charting indicated that the mean for homework completion for her class was 90% per week. The teacher also indicated that only 2 students were performing below the 25<sup>th</sup> percentile, 6 students were performing between the 25<sup>th</sup> and 50<sup>th</sup> percentiles, 8 students were performing between the 50<sup>th</sup> and 75<sup>th</sup> percentiles, and 10 students were performing above the 75<sup>th</sup> percentile. Carlos was initially performing below the 25<sup>th</sup> percentile with only a 50% homework completion rate and after intervention was performing above the 75<sup>th</sup> percentile with a 99% homework completion rate. The graph below compares Carlos' desired goal (set by the teacher) of 90% to the actual end results of 100% for homework completion.

Carlos' Rate of Homework Completion



Treatment integrity data indicated that the intervention was implemented with 100% integrity during the 8 week period. The teacher, Carlos' mother, and the consultant found the intervention to be extremely effective for Carlos. The team found one limitation to be concern regarding removal of the pizza lunch reward. The team was concerned when the reward was removed that Carlos' homework completion rate would decline. However, the team decided to slowly remove the pizza lunch reward and continue to monitor his homework completion rate. The team decided to slowly increase the number of weeks of 100% homework completion before Carlos would earn the pizza reward. The team also decided that after increasing the number of weeks for 100% homework completion to slowly transition Carlos to differing types of positive reinforcements, such as emails from school staff members or his mother telling him how good he is doing by completing his homework.

R002588754