DOCUMENT RESUME

ED 475 362 TM 034 838

AUTHOR	Giraud, Gerald T.; Buckendahl, Chad; Lucas, Mike
TITLE	Teacher Judgment and Mastery Score Setting for Nebraska's Local Assessment Reporting System: Efficacy and Appropriateness.
PUB DATE	2003-04-00
NOTE	17p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Classification; *Cutting Scores; *Elementary School Students; Intermediate Grades; *Mastery Tests; *Mathematics; *Student Evaluation
IDENTIFIERS	*Nebraska

ABSTRACT

This study examined the efficacy of teacher judgment in the process of setting mastery scores (cut scores) for fourth-grade mathematics in local school districts in Nebraska in terms of agreement between teacher classification of students and classification by the cut score obtained by this classification. The study also examined cut scores in terms of appropriateness by examining cut score differences across levels. Data, in the form of cut scores, student classification by teachers, and classification by cut scores, were obtained from 4 small-to-medium school districts for a total of 374 fourth graders. Analysis reveals that the classification of students into mastery levels is often no in agreement with student classification determined by the cut scores resulting from the modified contrasting groups method, a method dependent on teacher classification of students. The inherent limitations of the modified contrasting groups method as applied by the districts in this study suggest that other more rigorous standard setting methods should be applied. (Contains 11 figures and 5 references.) (SLD)

Teacher judgment and mastery score setting for Nebraska's local assessment reporting system: Efficacy and appropriateness

Gerald T. Giraud, Ph.D.

Nebraska Methodist College

Chad Buckendahl, Ph. D.

Buros Institute for Assessment Consultation and Outreach

Mike Lucas

West Point Community Schools

This paper is prepared for the:
Annual Meeting of the American Educational Research Association in Chicago, IL
April 2003

Introduction

Nebraska's approach to standards, assessment, and accountability - School-based Teacher-led Assessment and Reporting System (STARS)- is based on the premise that decisions about student learning should be standards-based and based upon teachers' knowledge of the student. This approach relies on the expertise of classroom teachers and their ability to assess and classify students as having met standards (established by the state department of education) in learning. Nebraska has therefore initiated an assessment system that relies on local school districts to design quality assessments and to make credible decisions about students' mastery of learning content based on these assessments. The state requires testing in reading, writing and mathematics at the 4th, 8th, and 11th grades. The setting of mastery (cut) scores is an important part of this assessment approach (Nebraska Department of Education, 2002b).

The Nebraska Department of Education (NDE) has developed a procedure for review of local assessments intended to assure the quality of locally developed assessments (NDE, 2002a, b). Districts are held accountable for 6 quality criteria, including match to standards, opportunity to learn, bias review, appropriate cognitive and developmental level of assessment, reliability, and appropriately determined mastery scores. It is the last of these quality criteria (mastery scores) that is the focus of this study. NDE describes the intent of this quality criterion as follows:

> This criterion is about determining "how good is good enough" in terms of levels of student achievement. Districts should provide evidence that the student mastery decisions were made using procedures that take into account the difficulty of the items or tasks in

> the assessments or classifications of students on an independent criterion. The procedure used to set mastery levels should include systematic judgments about assessment content and the difference levels of student performance. The important thing here is for districts to identify and describe the method used to set mastery levels. Districts should not rely on their traditional grading scale to make these decisions. Professional judgment needs to be used about students or about the test /work itself to arrive at mastery level decisions. (NDE, 2002a, page 14).

The current study focuses on one frequently employed method for setting mastery scores (cut scores) for 4th grade mathematics in local school districts in Nebraska.

NDE established standards in 6 areas of content for 4th grade mathematics (Numeration, Computation, Measurement, Geometry, Data Analysis, and Algebraic Concepts; NDE, 2002c). Within these 6 main content areas were a total of 18 more specific standards. The state requires that districts report the percentage of students attaining mastery in each of these 18 specific standards. Although NDE requires that districts report mastery or non mastery, many districts choose to categorize students into 4 levels of mastery: beginning, progressing, proficient, or advanced. The reason for this is reporting requirements for special education programs.

NDE suggested several strategies for setting cut scores on local assessments designed to determine student accomplishment relative to state standards (NDE, 2002a). Among these are familiar methods such as Angoff (1971) and its several modifications and a modified contrasting groups method. Many districts chose to use the later, because of its relative simplicity. In this method, as described by NDE, teachers categorize students by level of mastery of mathematics standards. After assessment, mathematics standards assessment scores of students in each category (as assigned by teachers) are averaged. Cut scores between categories are the average of adjacent category averages.

The modified contrasting groups method relies on teacher ability to correctly classify students in terms of mastery, and also on an assessment score's utility in reflecting mastery of standards content. Districts employed several strategies in application of this method. Some districts asked teachers to classify students into mastery levels on all 18 specific standards, others asked teachers to classify students on the six major standards, and still others simply asked teachers to classify students by their overall math ability. The first strategy requires teachers to make 18 mastery judgments per student, the second 6 and last of course only 1. Most districts included in this study classified students into 4 levels of mastery, as described above. No matter which strategy was employed, districts were required to set mastery scores (cut scores) for all 18 specific standards, and to report the percentage of students attaining mastery.

The purpose of this study is to examine the efficacy of teacher judgment in the process, in terms of agreement between teacher classification of students and classification by the cut score obtained by this classification. Further, this study examines the cut scores in terms of appropriateness, by examining cut score differences across levels. For example, are cut scores stepped as expected (that is, are cuts for lower levels of mastery lower than for higher levels?), and are cuts substantially different (e.g. how many points separate low and high levels of mastery?).

Method

School districts represented in this study used a modification of the contrasting groups strategy described by Livingston and Zieky (1982) as the method to set the passing scores for 4th grade mathematics assessments. In the typical application of this method, teachers classify students with whom they are familiar into four levels of mathematics proficiency (based on Nebraska state content standards) after a discussion of the characteristics of students in the four proficiency categories (beginning, progressing, proficient, advanced). Those classifications are then replaced with actual student performance to determine a recommended cut score. The cut score that separates categories is the mean or median of the mean or median of adjacent categories. For example, the cut score between beginning and progressing is the mean of the mean score of students classified as beginning and the mean score of students classified as progressing.

Ideally, the classification of students by teachers upon which the cut scores are eventually determined follows a thorough discussion of the standards and the characteristics of students in each performance category. However, these procedures are unstandardized, and the exact procedures followed by some of the districts represented in this study are unknown. It is known that the cut scores in each district were determined by teacher classification of students as described above.

Data

For this preliminary study, data in the form of cut scores, student classification by teachers and classification by cut scores were obtained from 4 small (300 or fewer students) to medium (300-1000) school districts in Nebraska and from a consortium of small to medium districts that shared the same assessments and worked together to set cut scores. A psychometrician expert in setting cut scores facilitated the process followed by the consortium. The other districts represented in the study completed the cut score setting procedures independently.

The total number of 4th grade students included in the study data is 374. Cut scores separating students into 4 categories (beginning, progressing, proficient and advanced) were determined for all 18 4th grade standards in each district, but the current study examines only data from 1 standard addressing numeration and number sense.

Analysis

Efficacy of teacher judgment in the modified contrasting groups method for setting cut scores as employed by school districts in Nebraska is examined by comparing classification by teachers to classification by cut score.

Appropriateness of cut scores is examined by comparing cut scores across levels, to determine whether there are substantive differences between cut scores that classify students into levels of proficiency.

Results

To determine the efficacy of teacher judgments, teacher classification of students was compared to classification based on student performance relative to the resulting cut scores. Table 1 shows that teacher and cut score classification disagree most often in the classification of students as progressing and proficient. In these categories, disagreement between teacher and cut score classification most often resulted in students being placed in higher categories by the cut score than by teacher rating. Agreement between teacher and cut score classification was low (Kappa=.233, se = .03, maximum .58).

Table 1. Teacher classification * Cut score classification Crosstabulation

			Cut score classification				
			Beginning	Progressing	Proficient	Advanced	Total
Teacher classification	Beginning	Count	20	8	2	3	33
		% within Teacher classification	60.6%	24.2%	6.1%	9.1%	100.0%
	Progressing	Count	27	19	27	27	100
		% within Teacher classification	27.0%	19.0%	27.0%	27.0%	100.0%
	Proficient	Count	16	13	46	74	149
		% within Teacher classification	10.7%	8.7%	30.9%	49.7%	100.0%
	Advanced	Count		3	13	76	92
		% within Teacher classification		3.3%	14.1%	82.6%	100.0%
Total		Count	63	43	88	180	374
		% within Teacher classification	16.8%	11.5%	23.5%	48.1%	100.0%

Because the assessments for which cut scores were set were different for each district, each district assessment was comprised of a different number of items, and the modified contrasting groups method might have been procedurally different in each district, an examination of the agreement between teacher classification and cut score classification by district is of interest. Table 2 reports results of agreement analysis by district. Agreement levels by district were comparably low.

Table 2. Agreement between teacher and cut score classification of students evaluated with kappa.

number of items on district assessment	Kappa	Asymp. Std. Error
16	.269	.108
26	.216	.063
50	.375	.106
56	.217	.086
31	.151	.043

The comments of an elementary principal, responsible for facilitating the cut score setting process in his school, reflected these results:

> At WP Elementary School, we were only successful about 55% of the time predicting the achievement category that our 39 students would achieve in this past spring when we administered our 4th Grade Math State Standards Assessment. We made predictions per each of the 18 standards because we felt that there could be some students that would be 4's or "advanced" in addition and subtraction but might only be 3's or "proficient" in multiplication. Our two 4th grade teachers, two 3rd grade teachers, and two other K-6 faculty members were in on the prediction process. These professionals used student report card grades, Terra Nova results from 2nd and 3rd grade, local CRT data, and attitude/effort as factors to consider when predicting levels of achievement. We consistently under-predicted scores for our highest achievers. For example, we too often predicted a 3 when students would easily achieve in the 4 range. I feel this will be a common find around the state. We also struggled a little in underestimating our low

achievers. There were several times we predicted a 1 when those students achieved in the 2, or even 3 categories. (Data report narrative from elementary principal, small district.)

The four-category classification strategy can also be collapsed to a two-category classification that identifies students as masters or non-masters of the content domain. In order to compare teacher and cut score classification for two categories, those students classified as beginning or progressing in either method were classified as non masters, and those classified as proficient and advanced were classified as masters. Agreement between teacher and cut score classification was improved when only two categories were used (Kappa = .444, se=. 05, maximum .835). Two-category comparison is shown in Table 3.

Table 3. Teacher classification * Cut score classification for Master and Non master.

			Cut score classification		
			Non master	Master	Total
Teacher classification	Non master	Count	74	59	133
		% within Teacher classification	55.6%	44.4%	100.0%
	Master	Count	32	209	241
		% within Teacher classification	13.3%	86.7%	100.0%
Total		Count	106	268	374
		% within Teacher classification	28.3%	71.7%	100.0%

Results of evaluation of agreement by district for two-category classification are reported in Table 4.

Tabel 4. Agreement between teacher and cut score classification for Master and Non Master evaluated by Kappa.

District	Kappa	Asymp. Std. Error
1	.499	.131
2	.403	.100
3	.674	.132
4	.522	.132
Consortium	.320	.084

Separation of category distribution is a necessary condition for the setting of appropriate cut scores. An examination of box plots reveals that teacher classification does result in some differentiation between categories of students, but that there is substantial overlap in some districts (See Figures 1-5).

Figure 1. District 1

16 items, KR21=.27

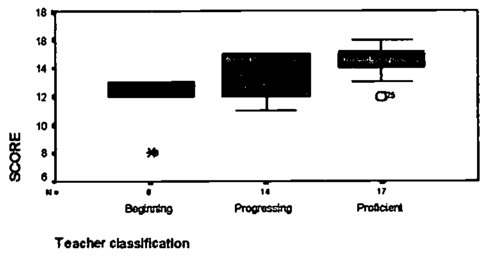


Figure 2. District 2

26 items, KR21=.78

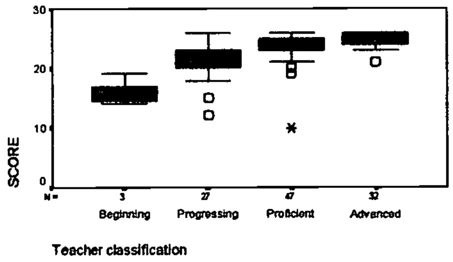


Figure 3. District 3

50 items, KR21=.96

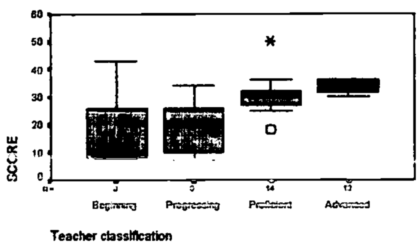


Figure 4. District 4

56 items, KR21=.91

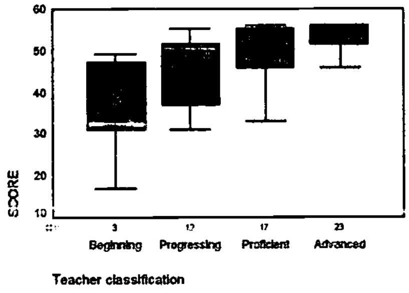
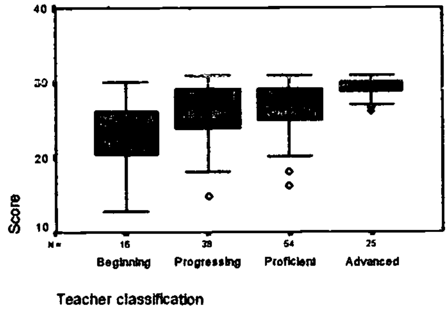


Figure 5. Consortium

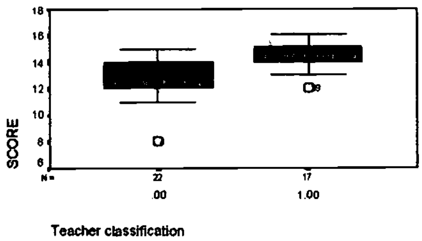
31 items, KR21=.74



An examination of box plots for two categories of student ability, based on teacher classification, reveals improved separation of distributions, but still substantial overlap (See Figures 6-10).

Figure 6. District 1

16 items



Firgure 7. District 2

26 items

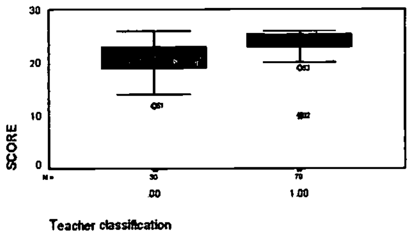


Figure 8. District 3

50 items

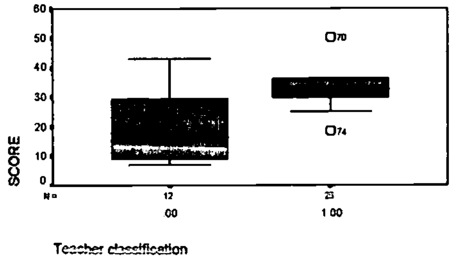


Figure 9. District 4

56 items

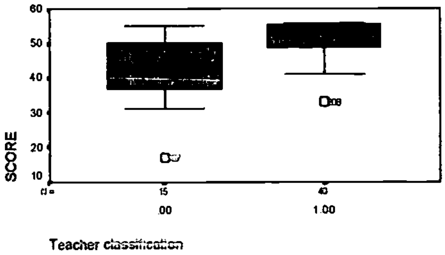
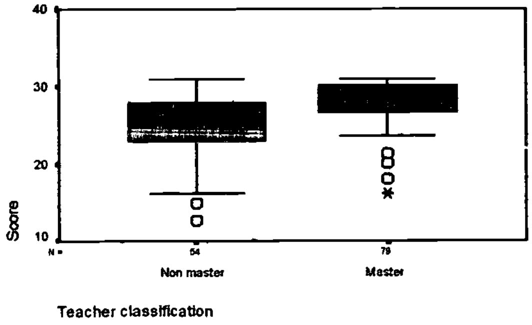


Figure 10. Consortium

31 items

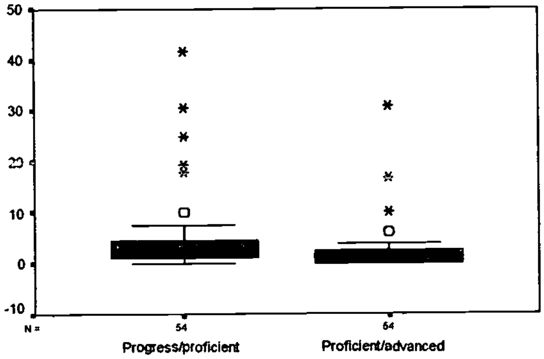


Appropriateness of cut scores determined by the modified contrasting groups method was examined by comparing cut scores across levels, to determine whether there were substantive differences between cut scores that classify students into levels of proficiency.

Two districts and the consortium provided cut scores for all 18 4th grade math standards assessments. This yielded a total of 54 sets of cut scores, and 162 individual cut scores (3 per standard). Two values are of interest: the difference between the cut score that categorized students as progressing and the cut score that categorized students as proficient, and the difference between the proficient and advanced cut scores. Thus, 54 values were derived for the former comparison, and 54 for the later.

Of the difference values for the progressing and proficient cuts, 54 % were 2 points or less and 70% were 3 points or less. For 6 comparisons (11%), the difference was 0. For the second comparison (proficient vs. advanced), 78% of differences were 2 points or less (64% 1 or 0); 90% were 3 or less. For 15 comparisons (28%), the difference was 0. See Figure 11. It should be noted here that when the method resulted in no difference between cut scores, one of two practices was followed: classification defaulted to master/non master, or a cut score above the calculated cut score for the higher category was adopted.

Figure 11. Cut score differences



Discussion

Analysis reveals that teacher classification of students into mastery levels is often not in agreement with student classification determined by the cut scores resulting from the modified contrasting groups method, a method dependent on teacher classification of students.
This lack of agreement could arise from teachers' inability to recognize the ability of their students to perform tasks specified by mathematics standards, from the inadequacy of the assessment used to measure student ability relative to the standards, from some aspect of the method used to determine cut scores, or from some combination of these factors.

This study focused on 4th grade standards, and therefor on 4th grade teachers. Elementary teachers in the districts represented in this study typically are with students most hours of every school day. Therefor, they should have the opportunity to know students and their ability, perhaps more so than, for example, high school teachers have the opportunity to observe and know the ability of older students. However, it is uncertain whether teachers know students well enough to make accurate categorizations of every student into one of 4 levels across 18 different (although related) mathematics standards. This remains to be studied.

The assessments that are intended to measure student ability relative to standards are locally developed in the Nebraska system, and this development is in its infancy. Based on this writer's experience with districts' efforts at assessment development, there is room for growth in terms of assessment quality. The assessment characteristic most likely to impede the accurate classification of students into several mastery levels is range of item difficulty. There is some evidence in this study that perhaps assessment items are not diverse enough in difficulty to cleanly separate advanced students from proficient students. For example, most (although not all) misclassification arises from students being classified lower by the teacher than by the assessment.

The modified contrasting groups method might also contribute to the mismatch between teacher and cut score classification of students. Using the mean of the mean of two adjacent categories when there is substantial overlap of score distributions suggests that the cut score will be near the middle of the distribution of scores within a classification level based on teacher judgment, thus assuring some level of discrepancy between teacher judgment and cut score classification. Figures 1-10 illustrate that this effect is not dependent on number of items, and is thus likely attributable to: a) assessments that cannot yield scores that separate students of differing ability, or b) the inability of teachers to correctly classify students.

The inherent limitations of the modified contrasting groups method as applied by the districts in this study suggest that other more rigorous standard setting methods be applied. For example, sample free methods like the Angoff (1971) that focus on item difficulty rather than examinee classification might be more defensible (if not more effective given the characteristics of assessments described above). If an examinee centered method is desirable because of resource limitations, then perhaps the borderline method, in which teachers identify students on the borders of mastery categories as a basis for cut score computation would result in cut scores more in agreement with teacher judgment about the ability of students.

This study suggests that Nebraska's plan for locally developed assessment for accountability and school improvement has some distance to go in establishing assessment quality, particularly in terms of setting defensible and sensible cut scores. It seems logical that the first step is to develop local assessments that have the psychometric characteristics necessary for separating students into several levels of ability. Further, consideration should be given to reducing either the number of standards for which cut scores are required, or the number of categories to be determined, in order to increase the likelihood that teachers will be able to accurately classify students, and that assessment results and cut scores will be useful and defensible indicators of ability. Finally, the modified contrasting groups method should be reconsidered as an acceptable method for determining cut scores.

References

Angoff, W. (1971). Scales, Norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington, D.C., American Council of Education.

Livingston, S.A., and Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princton, N.J.: ETS.

Nebraska Department of Education (2002). *Suggestions on preparing a district assessment portfolio describing six quality criteria for the 2001-2002 mathematics assessment.* Retrieved March 1, 2003, www.nde.state.ne.us/stars/HelpfulSuggestionsDAP.pdf.

Nebraska Department of Education (2002). *School-based Teacher-lead Assessment and Reporting System: A summary.* Retrieved March 30, 2003, www.nde.state.ne.us/stars/pdf/STARSbooklet.pdf.

Nebraska Department of Education (2002). *Academic Standards: Mathematics.* Retrieved March 1, 2003, www.nde.state.ne.us/LEGAL/Appendix%20A-clean.pdf.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title;
Teacher judgment and mastery score setting for Nebraska's local assessment reporting system; Efficacy and appropriateness

Author(s): G. Giraud, PhD; C. Buckendahl, PhD; M. Lucas

Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☒	Level 2A □	Level 2B □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and
its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature:	Printed Name/Position/Title: Gerald Giraud, Associate Professor	
Organization/Address: Nebraska Methodist College Omaha, NE	Telephone:	FAX:

	E-Mail Address: ggiraud@methodistcollege.edu	Date:

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200