ED 448 195	TM 032 225
AUTHOR	Buckendahl, Chad W.; Impara, James C.; Plake, Barbara S.
TITLE	Computing Composite Scale Scores for Accountability: A Validation Study of Nebraska's District Evaluation Model.
PUB DATE	2000-10-00
NOTE	15p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 25-28, 2000).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Accountability; *Classification; Comparative Analysis; Elementary Secondary Education; Evaluation Methods; Performance Factors; *Scaling; School Districts; *School Effectiveness; State Programs; Testing Programs; Validity
IDENTIFIERS	*Composite Scores; *Nebraska

ABSTRACT

Because districts in Nebraska are not measured on common instruments, comparisons are difficult. This study examined a district evaluation strategy that classifies districts into school performance ratings (SPR) based on a combination of three factors. Student performance and non-cognitive indicator data for three grade levels and two content areas from school districts in a southern state were used. Analyses comparing classification decision consistency for three mathematical models were conducted using Spearman rank order correlations for composite scale scores (CSS) and kappa statistics for SPR classifications. Results show that there is a high level of agreement among the three mathematical models considered, suggesting that the preferred model be the one that is easiest to understand and communicate. (Contains 2 tables and 10 references.) (Author/SLD)

Computing composite scale scores for accountability: A validation study of Nebraska's district evaluation model

Chad W. Buckendahl

James C. Impara

Barbara S. Plake

University of Nebraska – Lincoln

Paper presented at the annual meeting of the

Mid-Western Educational Research Association in Chicago, IL.

October, 2000

Abstract

Many states use a statewide assessment strategy to evaluate districts on common measures. Because districts in Nebraska are not measured on common instruments, comparisons are difficult. This study examined a district evaluation strategy that classifies districts into school performance ratings (SPR) based on a combination of three factors. Student performance and non-cognitive indicator data for three grade levels and two content areas from school districts in a Southern state were used. Analyses comparing classification decision consistency for three mathematical models were conducted using Spearman rank order correlations for composite scale scores (CSS) and kappa statistics for school performance ratings (SPR) classifications. Results show that there is a high level of agreement between the three mathematical models considered suggesting that the preferred model be the one that is easiest to understand and communicate.

Computing composite scale scores for accountability: A validation study of Nebraska's district evaluation model

Educational accountability is a common topic discussed among educators and administrators nationwide. It has become evident, though, that control over methods of accountability has shifted from the local jurisdiction (school districts) to the state jurisdiction (state departments of education and legislative entities). The shift is not surprising because popular media has given increasing attention to educational accountability. One reason may be the general belief that public education has not lived up to the expectations that have been placed on it.

Research on accountability systems is not new. Scholars have focused on outcome measures and the decisions that are generally associated with higher stakes accountability systems (Tyler, 1973; Dyer, Linn, & Patton, 1968). More recent research has addressed communicating meaningful results (Cornett & Gaines, 1997) and the feelings of other stakeholders in the system (King & Mathers, 1999; Law, 1999). The validity information provided by the selected measures in the system represents a critical element in evaluating district performance. Whether using norm-referenced or criterion-referenced instruments, the alignment and proper use of those instruments relative to the desired educational objectives is essential to an accountability system.

Additional problems arise when the scores are reported and schools' or districts' performance is compared. With a common assessment, states have rank ordered school districts based on performance at individual grades and content areas (e.g., Georgia) or on a composite index of district performance that considers non-cognitive indicators (e.g., Kentucky) or a composite index that does not include non-cognitive indicators (e.g., Texas). However, the rank ordering may not be meaningful without considering some of the non-cognitive indicators that may affect performance (Guskey & Kifer, 1990).

The purpose of this study was to evaluate the utility of composite scores calculated for a state accountability model in terms of classification decisions using real data from school districts. The goal was to compare the utility of three related mathematical models with each other. Variables for the study included: 1) estimates of student performance from district assessments and 2) ratings of the technical quality of district assessments. Third, variables that made adjustments to composite scores for districts that possessed student characteristics that presented high levels of challenge to district success were also included. These variables included proxies for socioeconomic status, students with disabilities, limited English proficiency, and inter-district mobility.

Methods

Two data sources were used for this study. The first source was student performance and district level non-cognitive indicator data from school districts in a Southern state. Districts in this state were selected because 1) the number of districts (67) in the state was considered reasonable and manageable as opposed to the number of districts in Nebraska (587), 2) a common measure of reading and mathematics achievement was available at grade levels (4th, 5th, 8th, and 10th) that were comparable to the Nebraska model, and 3) the non-cognitive indicators that the state collects included those of interest to Nebraska.

A second data source consisted the judgments of an educational advisory committee that was comprised of educational representatives from across the state of Nebraska. The committee provided judgments of appropriate weights for the three components (student performance, technical quality, and non-cognitive indicators) that form the composite scale score (CSS) for a district.

Information on the technical quality of assessment strategies was not available for the state selected because they use a common assessment for reading and mathematics across all districts. Therefore, for this study, technical quality ratings were randomly assigned to districts to simulate possible combinations that districts in Nebraska may exhibit in practice. The technical quality rating, which ranges from 1 to 5 was transformed to a scale that ranged from 0 to 1 to keep the resultant CSS on a predetermined scale when they were combined.

Judgments collected by the educational advisory committee were gathered over two rounds. After a brief training exercise, members provided judgments for round one. This initial round was followed by feedback data. The feedback data consisted of the mean weighting for each component as indicated by the group and also the range of judgments. Judges were then allowed to reconsider their initial judgments in round two. The results collected from the educational committee member judgments in round two were then used for the judgmental weights in forming the CSS.

Procedures

Empirical weights for the four non-cognitive indicators were determined through separate regression analyses for each grade level and content area using Florida

Comprehensive Achievement Test (FCAT) reading and mathematics scores as dependent variables. The analyses were conducted with the four non-cognitive indicators as predictor variables and used for each content area across the three grade levels. Each non-cognitive indicator was coded dichotomously (0 – below the state average and 1 – at or above the state average) when considered in the equation. This means that various combinations of non-cognitive indicators could be included into a district's equation conditional on their position on a specific non-cognitive indicator relative to other districts in the state.

Judgmental weights were determined after a meeting with the educational advisory committee using a multiattribute utility theory (MAUT) approach (Jaeger & Usher, 1991). Members of the advisory committee engaged in a brief training session in which the components of the CSS (student performance, student performance, non-cognitive indicators – socioeconomic status, students with disabilities, limited English proficiency, and mobility) were described in detail. The committee then provided two rounds of ratings of the relative contribution of the components. After the committee provided their first round judgments on the relative contribution of each of the components, feedback on the committee's judgments was presented to allow for reconsideration of their judgments in a second round of ratings.

The three mathematical models that were analyzed using this information were:

1. $(JW_1A \text{ x } JW_2B) + [EW_1C_1 + EW_2C_2 + EW_3C_3 + EW_4C_4]= CSS$ (Empirical Model)

2. $(JW_1A \text{ x } JW_2B) + [JW_3C_1 + JW_4C_2 + JW_5C_3 + JW_6C_4]= CSS$ (Judgmental Model)

3. $(A \text{ x } B) + [C_1 + C_2 + C_3 + C_4]= CSS$ (Unweighted Model)

In the models, "A" represents student performance, "B" represents technical quality, and "C" represents a non-cognitive indicator. For each of these models empirical weights (EW) were determined through regression analyses. Judgment weights (JW) were determined by judgments of members of the educational advisory committee.

To transform composite scale scores into the five school performance ratings (SPR) categories, a decision rule was created to specify the cut point for each level. The overall composite scale scores range from 1-100, however, for districts above the state average on non-cognitive indicator variables, it was possible for their scores to be as high as 110. The width of each range was chosen to represent a symmetrical distribution with an equivalent number of scores in the upper and lower ranges rather than a uniform distribution that had an equal number of values in each score interval.

Two analysis methods, Spearman rank order correlations and kappa statistics, were used to examine model classification decision consistency for each combination of grade level (4th, 5th, 8th, and 10th) and content area (reading and mathematics). Initially, Spearman rank order correlations (Siegel & Castellan, Jr., 1988) were conducted on the resultant composite scale score rank ordering of districts for each model to determine the level of agreement between the pairs of models considered.

However, because the final SPR classification is a rating system, kappa statistics (Siegel & Castellan, Jr., 1988; Traub, 1994) were calculated to measure the level of agreement between the pairs of model classification decisions of districts. Using the three models as classification instruments, a "step-up" analysis was performed that first compares the classification decision agreement between the least complex model considered and the second least complex model considered. The second "step" then compares classification decision agreement between the second least complex model considered and the most complex model considered.

Results

Correlations between non-cognitive indicators and student performance data used for the regression analyses used to determine empirical weights are reported in Table 1. As seen in this table, moderate negative correlations are seen for the socioeconomic status variable (SES) and test scores for all grades and content areas. This suggests that having a higher percentage of economically disadvantaged students will lower test scores. Limited English Proficiency (LEP) for 4th and 8th grade reading and Mobility (MOB) for 8th grade reading also correlated negatively, but at a lower magnitude than SES. The presence of students with disabilities (IEP) did not correlate significantly with any of the test score variables. It is interesting to note, however, that IEP and LEP were negatively correlated with each other. This may suggest that there are classification decisions that are made within school districts that impact these variables (i.e., when one increases, the other decreases).

A Spearman rank order correlation coefficient was calculated between the resultant CSS of the unweighted and judgmentally adjusted models for each of the combinations of grade level and content area. Values were as follows: 4th grade reading, $r = .999$ ($p < .0001$); 5th grade mathematics, $r = 1.000$ ($p < .0001$); 8th grade reading, $r = 1.000$ ($p < .0001$); 8th grade mathematics, $r = 1.000$ ($p < .0001$); 10th grade reading, $r = 1.000$ ($p < .0001$); and 10th grade mathematics, $r = 1.000$ ($p < .0001$). These values indicate a high relationship of the CSS between models. Because the component weights for the unweighted and the judgmentally adjusted models were very similar, these high correlations were expected.

A Spearman rank order correlation coefficient was also calculated between the resultant CSS of the judgmentally adjusted and empirically adjusted models for each of the combinations of grade level and content area. Values were as follows: 4th grade reading, r = .978 (p < .0001); 5th grade mathematics, r = .984 (p < .0001); 8th grade reading, r = .986 (p < .0001); 8th grade mathematics, r = .989 (p < .0001); 10th grade reading, r = .988 (p < .0001); and 10th grade mathematics, r = .988 (p < .0001). These values also indicate a high relationship of the CSS between models, yet slightly lower than the relationship between the first two models.

The calculated value for kappa represents that level of agreement adjusted for the possibility the agreement was by chance. Table 2 shows resultant kappa values for each grade level and content area. Statistical significance tests were also run to determine if the observed kappa value was beyond what was expected by chance. All kappa values were statistically significant at the .0001 level.

Conclusions and Implications

Of the three models considered in this study, the preferred model would likely be the unweighted model. Since the judgmental and unweighted models were essentially the same and the empirical model generally left a district's SPR classification unchanged relative the unweighted SPR, the unweighted model would likely be the easiest to understand. From a policy perspective, because the current system is low stakes, the slight loss in classification accuracy in selecting the unweighted model does not outweigh the need for simplicity of understanding for stakeholders.

Future research could address the choice of mathematical models and the scales that transform the CSS to the SPR. The models that were considered in this study multiplied student performance by technical quality and then made adjustments based on the number of non-cognitive indicators that a district was at or above the state average. Other models, such as a purely additive or purely interactive model could be considered. The cut points on the scale, made it very difficult for a district to achieve a SPR rating of "5". Using real data from school districts in this study, no districts among the 67 were classified as a "5". If additional external validity evidence suggests that classifications are underestimating district performance, a re-calibration of the scale may be warranted. Finally, it is important to remember that this accountability system was developed with no stakes (beyond public opinion) associated with a school district's performance. Replication of this study with a high stakes system may not be appropriate.

References

Cornett, L.M. & Gaines, G. (1997). Accountability in the 19990s: Holding Schools Responsible for Student Achievement. Atlanta, GA: Southern Regional Education Board.

Dyer, H.S., Linn, R.L, & Patton, M.J. (1968). Methods of Measuring School System Performance. Princeton, NJ: Educational Testing Service.

Florida Department of Education. (2000). Florida Department of Education Website. Website can be accessed at http://www.firn.edu/doe.

Guskey, T.R. & Kifer, E.W. (1990). Ranking school districts on the basis of statewide test results: Is it meaningful or misleading? Educational Measurement: Issues and Practice, 9(1), 11-16.

Jaeger, R.M. & Usher, C.H. (1991, April). Alternative procedures for integrating multidimensional evaluations of schools: An experimental comparison. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL.

King, R.A. & Mathers, J.K. (1999). Financing schools based on performance measures. School Business Affairs, 65(1), 3-9.

Law, N. (1999). Value-added assessment and accountability. Thrust for Educational Leadership, 28(3), 28-31.

Siegel, S. & Castellan, Jr., N.J. (1988). Nonparmetric Statistics for the Behavioral Sciences (2nd ed.) pp. 262-291. New York, NY: McGraw-Hill, Inc.

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

Tyler, R.W. (1973). Testing for Accountability. In R.W. Hostrop, J.A. Mecklenburger, & J.A. Wilson (Eds.), Accountability for Educational Results, pp. 159-162. Hamden, CT: Linnet Books.

Table 1.

Correlations between non-cognitive indicators and student performance data.

	SES	IEP	LEP	MOB	Read4	Math5	Read8	Math8	Read10	Math10
SES	1.000									
IEP	.151	1.000								
LEP	.004	-.256#	1.000							
MOB	.184	.169	.182	1.000						
Read4	-.693*	-.013	-.310#	-.121	1.000					
Math5	-.607*	-.062	.002	.067	.773*	1.000				
Read8	-.656*	.105	-.281#	-.243#	.782*	.645*	1.000			
Math8	-.682*	-.071	-.185	-.239	.778*	.713*	.917*	1.000		
Read10	-.693*	-.016	-.210	-.196	.697*	.477*	.760*	.751*	1.000	
Math10	-.699*	.017	-.120	-.157	.738*	.591*	.773*	.777*	.940*	1.000

* Correlation is significant at the .01 level (2-tailed)
# Correlation is significant at the .05 level (2-tailed)

Table 2.

Kappa statistics of pairwise model classification agreement.

	Unweighted – Judgmental	Judgmental - Empirical
Grade level and content area:		
4th Grade Reading	.980	.861
5th Grade Mathematics	1.000	.940
8th Grade Reading	1.000	.920
8th Grade Mathematics	1.000	.940
10th Grade Reading	1.000	.900
10th Grade Mathematics	1.000	.940
Mean Kappa	**.997**	**.917**

* All kappa values are statistically significant at the .0001 level.

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

ERIC®

TM032225

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Computing composite scale scores for accountability: A validation study of Nebraska's district evaluation model.

Author(s): Chad W. Buckendahl, James C. Impara, & Barbara S. Plake

Corporate Source: University of Nebraska - Lincoln

Publication Date: October, 2000

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

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

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

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

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: Chad W. Buckendahl

Printed Name/Position/Title: Chad W. Buckendahl, Ph.D.

Organization/Address: Buros Center for Testing
135 Bancroft Hall, UNL
Lincoln, NE 68588-0352

Telephone: (402) 472-6244

FAX: (402) 472-6207

E-Mail Address: biaco@unl.edu

Date: Dec. 11, 2000

(over)

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:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20772
ATTN: ACQUISITIONS

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
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com