DOCUMENT RESUME

ED 437 349 SP 038 929

AUTHOR	Mertler, Craig A.
TITLE	Teachers' (Mis)Conceptions of Classroom Test Validity and Reliability.
PUB DATE	1999-10-00
NOTE	18p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 13-16, 1999).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; Evaluation Methods; *Student Evaluation; Teacher Attitudes; Teachers; *Test Reliability; *Test Validity; Testing
IDENTIFIERS	Ohio

ABSTRACT

This study examined processes and techniques teachers used to ensure that their assessments were valid and reliable, noting the extent to which they engaged in these processes. A sample of 625 elementary and secondary teachers received mailed copies of the Ohio Teacher Assessment Practices Survey, which asked about steps that they followed and the extent to which they went to ensure that their assessments were valid and reliable. Results indicated that teachers did not spend much time conducting statistical analyses of their assessment data. Steps teachers used to ensure validity fell into six categories: using teacher-developed tests, comparing to objectives or curriculum, analyzing test data, not determining validity, asking for student feedback, and miscellaneous. Over half of the respondents used teacher-made tests. Most ensured assessment validity by following conventional rules of sound test development. Many believed that developing their own assessments would ensure validity. About 30 percent said that they followed specific steps to ensure reliability half of the time or less. Strategies to ensure reliability fell into six categories: using teacher-developed tests, comparing to objectives or curriculum, analyzing test data, using same process as validity, asking for student feedback, and miscellaneous. Many teachers believed that ensuring assessment validity and reliability were very similar processes. Teachers had a better grasp of the concept of reliability than of validity. (SM)

TEACHERS' (MIS)CONCEPTIONS OF CLASSROOM TEST VALIDITY AND RELIABILITY

Craig A. Mertler, Ph.D.
Educational Foundations & Inquiry Program
School of Educational Leadership & Policy Studies
College of Education & Human Development
Bowling Green State University
Bowling Green, OH 43403

A Paper Presented at the 1999 Annual Meeting of the
Mid-Western Educational Research Association

Chicago, Illinois

October 13-16, 1999

BACKGROUND

A sizable amount of classroom time is devoted to the assessment of student learning. Since teachers must give even more time to the preparation and scoring of tests and other assessments, a substantial proportion of a teacher's day is devoted to issues surrounding student assessment. One could argue, then, that careful consideration of testing within formal teacher preparation programs is certainly warranted. If educators, particularly those in teacher preparation programs, are to help teachers use their student testing time efficiently and to be effective at it, more must be learned about how teachers perceive and use classroom tests and other forms of assessment (Gullickson, 1984).

Several research studies examining the overall assessment practices of classroom teachers have been conducted; however, little research on the topic of practices with respect to insuring classroom test validity and reliability exist in the literature. Much of the research has focused on the use of various types of items and differences that exist across school levels (i.e., elementary, middle, and high schools) and school locations (i.e., urban, suburban, and rural). For example, Marso (1985; 1987) found several differences between elementary and secondary teachers. Secondary teachers tended to use more self-constructed tests rather than published tests; whereas, the opposite was true for elementary teachers, especially those in grades K-4. Similarly, others have found that the higher the grade level, the greater the tendency for teachers to use their own assessments (Stiggins & Bridgeford, 1985). Secondary teachers reported relatively more use of essay and problem-type items and less frequent use of completion and multiple-choice items than did elementary teachers (Marso & Pigge, 1987). Marso (1985) also found that teachers perceived matching, multiple-choice, and completion type items as being most useful.

This research was made possible by a grant from the Martha Holden Jennings Foundation, Cleveland, OH (Grant No. A-107-97). The author wishes to express his appreciation to the Foundation for its support.

Very little research exists on how teachers determine the extent to which their assessments are valid and reliable. However, some research on teachers' use of statistical analyses of test data does exist. A final overriding theme in studies of teachers' assessment practices is the infrequent use of statistical analyses of test data (Gullickson, 1986; Marso & Pigge, 1987; Marso & Pigge, 1988). This may be due to the fact that teachers are not convinced of the value of using statistical procedures to improve the quality of their tests or that they simply do not have a good grasp of statistical concepts and this discomfort may lead to a devaluing of their use.

This study was part of a larger research endeavor which had as its main purpose the examination of the current assessment practices of K-12 teachers in the state of Ohio. The researcher sought to explore how practicing teachers assess student performance with their students in their own classroom settings. Specifically, the goal of this research study was to gain an understanding of the processes and techniques used by classroom teachers to insure that their assessments are both valid and reliable, and to determine the extent to which they engage in these processes.

METHODOLOGY

The researcher made use of resources available through the Ohio Department of Education in order to obtain a stratified random sample of K-12 teachers throughout the state of Ohio. The sample was stratified so that various subgroups in the population of K-12 teachers in the state were represented in the sample in the same proportion that they exist in the population. These subgroups of teachers included the following four categories: (1) female elementary, (2) female secondary, (3) male elementary, and (4) male secondary. A random sample of 3,000 teachers was obtained.

An original survey instrument, the *Ohio Teacher Assessment Practices Survey*, was developed by the researcher for purposes of collecting the data. The literature was relied upon heavily in order to guide the development of the specific items appearing in the survey instrument. The instrument consisted of 47 items and included both scaled (forced-choice) and open-ended items. For purposes of the study at hand, teachers were asked to respond to items concerning the validity and reliability of their classroom assessments, specifically requesting information on the steps that they follow and the extent to which they do so.

In mid-January, each teacher received a packet containing a full-page cover letter, copy of the survey, and a self-addressed, postage-paid return envelope. They were instructed to return the survey within four weeks from the date appearing on the cover letter. In mid-February, a follow-up reminder postcard was sent to those teachers who had not yet returned completed surveys. The final sample upon which the analyses were conducted consisted of 625 completed surveys. Analyses were conducted using SPSS (v. 6.1) and NUD*IST (v. 4).

RESULTS

The sample consisted of 53% females and 47% males. The majority (42%) of teachers were from suburban settings, followed closely by rural (32%) and urban (25%). Nearly half (47%) were teaching at the senior high level; just over one-fourth (26%) were teaching at the elementary level, followed closely by those teaching at the junior high/middle school level (25%). Twenty percent of the teachers had 26-30 years of teaching experience, followed by 21-25 years (19%), 6-10 years (17%), 1-5 years (13%), 16-20 years (13%), 11-15 years (11%), and 31-35 years (6%). Two teachers in the sample had 36 years or more of teaching experience.

Validity of Classroom Assessments

Teachers were asked to list specific steps they followed to insure that their assessments were valid and to indicate how often they followed these steps. One-fourth

(25%) of the teachers responded that they followed specific steps to insure validity about half of the time or less; the median response was "most of the time."

When asked to provide the specific steps that they follow to insure validity, the teachers provided a wide variety of responses. Six hundred and eleven responses were examined and categorized based on common approaches. The resulting hierarchical coding system is shown in Figure 1.

Insert Figure 1 about here

The teachers' responses were coded into six major categories, with the vast majority falling into roughly two of those categories. The major categories, with the numbers and percentages of response appearing in parentheses, were as follows:

- teacher-developed tests (352 or 58%);
- compare to objectives or curriculum (110 or 18%);
- analysis of test data (54 or 9%);
- don't determine validity (41 or 7%);
- ask for student feedback (20 or 3%); and
- miscellaneous (27 or 4%);

Several of these major categories included anywhere from a couple to several sub-categories. The sub-categories, along with the frequencies and percentages of response, are provided in Figure 1.

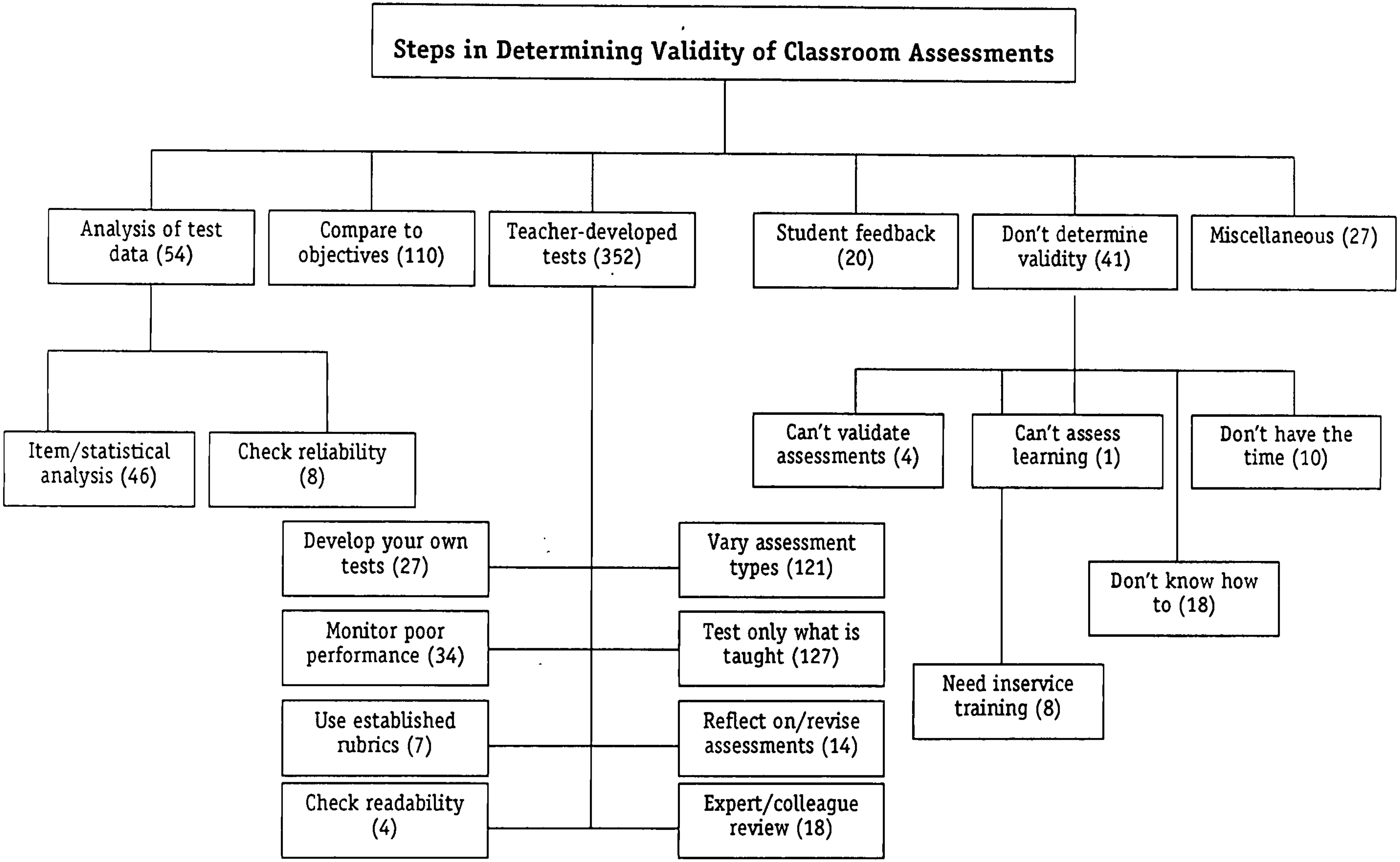
Steps in Determining Validity of Classroom Assessments
Analysis of test data (54)
Compare to objectives (110)
Teacher-developed tests (352)
Student feedback (20)
Don't determine validity (41)
Miscellaneous (27)
Item/statistical analysis (46)
Check reliability (8)
Can't validate assessments (4)
Can't assess learning (1)
Don't have the time (10)
Develop your own tests (27)
Vary assessment types (121)
Don't know how to (18)
Monitor poor performance (34)
Test only what is taught (127)
Need inservice training (8)
Use established rubrics (7)
Reflect on/revise assessments (14)
Check readability (4)
Expert/colleague review (18)

Figure 1. Coding scheme for teachers' approaches to determining classroom assessment validity

As evident in Figure 1, more than half of the responses dealt with teacher-made tests. The vast majority of teachers stated that they insure assessment validity by following conventional rules of sound test development, varying the types of items and assessments (thus providing students different means of showing what they know), and by simply testing what is taught. Several comments exemplifying these points follow:

> *Vary the type of questions in terms of difficulty and questioning types.*
>
> *Use essay questions, not multiple guess or True or False...I did not know what students understood giving multiple choice and True or False tests. Essays are more time consuming to grade, but well worth the effort.*
>
> *Make sure all tests are varied enough in questioning to accommodate all learning styles, not just one or two styles of questions.*
>
> *Written tests should be based entirely on what was taught.*

Many teachers believe that simply developing your own assessments, as opposed to using published materials, will insure assessment validity. Other teachers tend to monitor their students' performance on the self-developed assessments; if students perform poorly, they make adjustments accordingly. One teacher stated

> *I take the tests as I go. If there are questions that most students bomb, I'll eliminate it, but if I feel they were well prepared for it, I'll keep the question.*

Several teachers believe that simply reflecting on the success of an assessment instrument, evaluating how well students performed, and then revising the instrument would insure validity. Along these lines, teachers suggested asking questions of the students in order to gather feedback concerning the assessment. For example,

> *I...have them evaluate the test according to what I taught or thought I taught.*

Finally, with respect to teacher-made assessments, a small sampling stated that they have "experts" or other teachers review their tests and other assessments as a means of checking the validity:

> *My colleagues and I pass tests around to each other to see if everyone is on the same level.*

Many teachers insure validity by comparing their assessments to instructional objectives or the district/statewide curriculum.

> *Compare assessment to objectives in order to evaluate individual questions.*
>
> *Ask questions based on the material to be learned/course of study/curriculum...try to see what they know as well as what they don't know.*

Many teachers rely on the results of statistical analyses of test data or other information resulting from assessments. Several teachers stated that they simply "checked reliability" as a means of insuring validity, without providing any details of how they did so. Others specifically stated that they conducted item analyses of student data, although their approaches to doing so may have been a little vague:

> *Use statistics to validate the reliability.*

A small proportion of teachers stated that they didn't attempt to validate their assessments for a variety of reasons including the fact that validation cannot be done, student learning cannot truly be assessed, and there just is not enough time to do so. However, the majority of teachers who responded in this category confessed that they simply did not know how to validate their assessments and that inservice training was desperately needed.

> *Get professionals to inservice with applications for practical use. Experiment with these methods. Choose the methods which best fit the specific needs.*
>
> *Teachers need concrete examples and explicit instruction on how to create valid assessment items for written tests.*

Miscellaneous comments covered a wide range and encompassed several areas not covered by the broad categories. These included comments related to comparisons to proficiency test scores, the issue of cheating, and taking the test yourself to see if it appears valid.

It is clear that, although many of these comments provide sound advice for teachers to follow, these "steps" simply are not appropriate for determining the validity of classroom assessments. By following good test development guidelines, teachers will certainly be more likely to achieve tests that are valid, but simply following those rules will not insure validity. Many teachers seemed to have the concept of reliability confused with that of validity when they identified item analyses as a means of validation.

For many classroom assessments, content validity would be the most important type of validity to establish. Unfortunately, only a few teachers mentioned the idea of an expert or outside review of an assessment instrument or activity. This is the widely accepted method of determining content validity of any type of instrument (Gay & Airasian, 2000). Careful planning will also help with assessment validity, but would require more than a simple comparison to objectives.

It should be noted that several teachers provided miscellaneous comments that definitely could not be considered means of establishing validity and appeared to be somewhat troublesome. These included:

> *Although my techniques are not written down any longer, I use a mental format which I change as needed. Experience is a wonderful resource.*
>
> *Over the years, you'll find out what works for you.*
>
> *It takes me over an hour to even write a new test. To be honest, other than using my experience, I don't have much time to worry about how valid my test is.*
>
> *I don't know. Most of the time I am so busy I don't have time to check validity. I guess I leave this job up to someone else.*
>
> *No clue! I have no training is doing this, and never really thought about it until reading this question.*
>
> *Teachers don't have time for this type of analysis! Why don't you teach in a public school for a year and find out what it is really like.*

Reliability of Classroom Assessments

Teachers were also asked to list specific steps they followed to insure that their assessments were reliable and to indicate how often they followed these steps. Nearly one-third (30%) of the teachers responded that they followed specific steps to insure reliability about half of the time or less; the median response was "most of the time."

When asked to provide the specific steps that they follow to insure reliability, the teachers again provided a wide variety of responses. Four hundred and thirty-one responses were examined and categorized based on common approaches. The resulting coding system is shown in Figure 2.

Insert Figure 2 about here

The teachers' responses were coded into five major categories, with the vast majority falling into one of those categories. The major categories, with the numbers and percentages of response appearing in parentheses, were as follows:

- teacher-developed tests (234 or 54%);
- compare to objectives or curriculum (59 or 14%);
- analysis of test data (47 or 11%);
- same process as validity (25 or 6%);
- ask for student feedback (21 or 5%); and
- miscellaneous (37 or 9%);

Several of these major categories included anywhere from a couple to several sub-categories. The sub-categories, along with the frequencies and percentages of response, are provided in Figure 2.

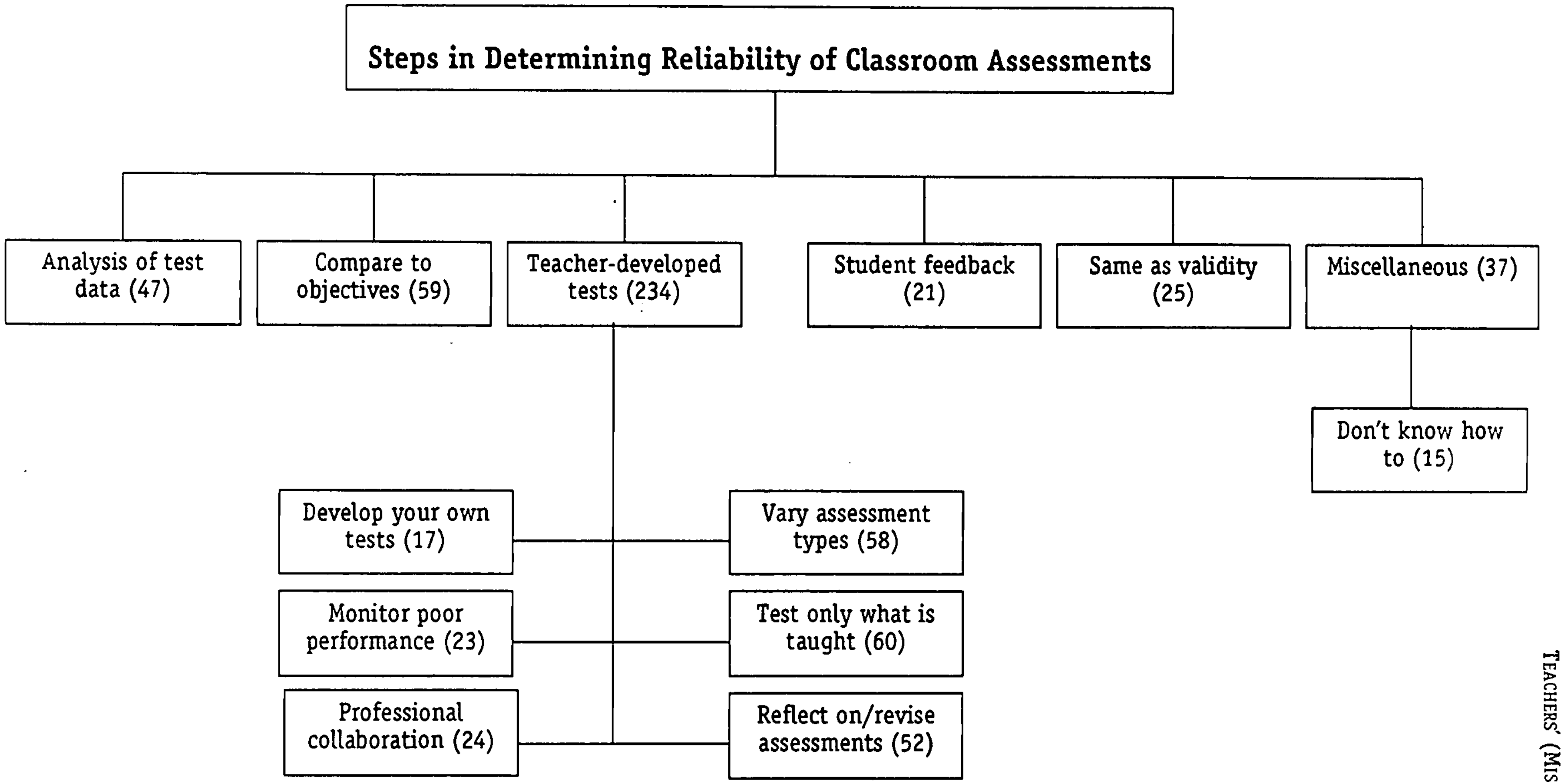
Steps in Determining Reliability of Classroom Assessments
Analysis of test data (47)
Compare to objectives (59)
Teacher-developed tests (234)
Student feedback (21)
Same as validity (25)
Miscellaneous (37)
Don't know how to (15)
Develop your own tests (17)
Vary assessment types (58)
Monitor poor performance (23)
Test only what is taught (60)
Professional collaboration (24)
Reflect on/revise assessments (52)

Figure 2. Coding scheme for teachers' approaches to determining classroom assessment reliability

As is evident from Figure 2, many teachers belief that insuring assessment validity and reliability are very similar procedures. Many of the same coding categories emerged as a result of examination of the responses to question addressing classroom assessment reliability. Again, the majority of teachers stated that they insure assessment reliability by following conventional rules of sound test development, varying the types of items and assessments, and by simply testing what is taught.

Similar to the responses regarding validity, teachers tend to monitor their students' performance on the self-developed assessments and revise them accordingly, as well as gathering oral feedback from students themselves about the assessment instruments or activities.

Again, many teachers identified professional collaboration as a means of insuring reliability, as well as comparing assessments to instructional objectives. Many teachers rely on the results of statistical analyses of test data or other data resulting from assessments to insure reliability. Several teachers explicitly stated that they utilized "test-retest" or "equivalent forms" methods of determining the extent to which their assessments are reliable.

A small proportion of teachers again stated that they did not know how to demonstrate the reliability of their assessments and that inservice training was necessary.

Miscellaneous comments included those related to comparisons to proficiency test scores, a teacher's knowledge of the content, performing readability tests on assessment instruments, and establishing a consistent grading system.

It seems that many teachers have a better grasp of the concept of reliability that validity, especially in terms of establishing those characteristics for their classroom assessments. However, the overriding majority of comments provided would not be considered acceptable methods of determining classroom assessment reliability.

It should be noted that several teachers provided miscellaneous comments that should again "raise a red flag" concerning their knowledge and ability to appropriately assess reliability. These included:

> *...techniques such as test-retest are possible, but they aren't practical in day to day classroom.*
>
> *Check grades...compare scores with what was taught. Use common sense.*
>
> *I would determine the percentage of students who demonstrate the ability you're looking for. Determine a ranking (90% answer correctly, then it is reliable).*
>
> *No specific steps. There are too many other things required of teachers.*
>
> *I really don't understand the difference between validity and reliability...sorry! Is it just me?*
>
> *...with all the other tasks at hand, worrying about the reliability of my tests is way down at the bottom of my priority list. I use my experience to determine reliability...*
>
> *What's the difference between reliable and valid -- really?*

CONCLUSIONS

This study was part of a larger research endeavor which had as its main purpose the examination of the current assessment practices of K-12 teachers in the state of Ohio. Specifically, the goal of this research study was to gain an understanding of the processes and techniques used by classroom teachers to insure that their assessments are both valid and reliable, and to determine the extent to which they engage in these processes.

This study was successful in that it resulted in a somewhat thorough description of these teachers' assessment practices with respect to issues of validity and reliability of their classroom assessments. It builds on previous classroom assessment practices research by incorporating information about validity and reliability analyses, which is quite scarce. Similar to previous research, it was determined that teachers do not spend much time conducting statistical analyses of their assessment data.

The results of this study perhaps imply that some attention needs to be re-focused on undergraduate teacher preparation measurement courses, especially in the areas of validity and reliability. Although these teachers *claim* they do a good job of following steps to insure sound assessments, they do not possess a solid foundation of what those steps should be. In other words, they frequently evaluate validity and reliability, but do so in the wrong ways. Therefore, they are really not evaluating those critical characteristics of classroom assessments. Only when measurement courses provide solid foundational understanding of these concepts will we have adequately prepared our teachers to assess their students' performance.

However, it may be more appropriate to focus teaching and training efforts on inservice—rather that preservice—teachers. It may be that teachers in general need to and should have some teaching and assessment experience—beyond the training received during student teaching—prior to being able to completely understand the concepts of validity and reliability, be able to consider those concepts during the development of their classroom assessments, and be able to appropriately assess these characteristics. Professional development is definitely something that teachers in this study identified as being necessary, needed, and useful.

REFERENCES

Gay, L.R. & Airasian, P. (2000). *Educational research: Competencies for analysis and application (6th ed.)*. Upper Saddle River, NJ: Merrill.

Gullickson, A.R. (1984). Teacher perspectives of their instructional use of tests. *Journal of Educational Research, 77*(4), 244-248.

Gullickson, A.R. (1986). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement, 23*(4), 347-354.

Marso, R.N. (1985, October). *Testing practices and test item preferences of classroom teachers*. Paper presented at the annual meeting of the Mid-Western Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 268 145)

Marso, R.N. & Pigge, F.L. (1987, October). *Teacher-made tests: Testing practices, cognitive demands, and item construction*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. (ERIC Document Reproduction Service No. ED 298 174)

Marso, R.N. & Pigge, F.L. (1988, October). *An analysis of teacher-made tests and testing: Classroom resources, guidelines, and practices*. Paper presented at the annual meeting of the Mid-Western Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 291 781)

Stiggins, R.J. & Bridgeford, N.J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*(4), 271-286.

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

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Teachers' (Mis)Conceptions of Classroom Test Validity and Reliability

Author(s): Craig A. Mertler, Ph.D.

Corporate Source:

Publication Date: October, 1999

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: Craig A. Mertler

Printed Name/Position/Title: Dr. Craig A. Mertler, Asst. Professor

Organization/Address:
EDFI Program
Bowling Green State University
~~Bowling Green, OH 43403~~

Telephone: 419-372-9357

FAX: 419-372-8265

E-Mail Address: mertler@bgnet.bgsu.edu

Date: 10/28/99

(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: THE ERIC CLEARINGHOUSE ON TEACHING
AND TEACHER EDUCATION
ONE DUPONT CIRCLE, SUITE 610
WASHINGTON, DC 20036-1186
(202) 293-2450

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
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

PREVIOUS VERSIONS OF THIS FORM ARE OBSOLETE.