

# The National Teaching & Learning FORUM



Volume 7

Number 6

1998

## CONTENTS

- **EDITOR'S NOTE:** *Aftermath*, p. 3. Learning from mistakes.
- **Making the Most of Exams: Procedures for Item Analysis**, Raymond M. Zurawski, St. Norbert College, p. 1. Was it a good test? How can you tell?
- **DEVELOPER'S DIARY:** *The Angstful Professor*, Laura Border, University of Colorado - Boulder, p. 5. That one dissenting vote often has the most to teach us about our teaching.
- **INNOVATIONS:** *A University Webzine to Promote Teaching and Learning*, Jan Smith and Paul Baepler, University of Minnesota-Twin Cities, p. 6. Minnesota invented the Gopher; they're still leading the way in serving academe via the Web.
- **VIEWPOINT:** *Love and Its Place in Mathematics*, David S. Tartakoff, University of Illinois at Chicago, p. 8. Will "wrong" answers die off like dinosaurs if the fear of learning wanes?
- **COUNTERPOINT:** *Wrong Answers?* Practical ideas in a more conventional approach from the Teaching Resource Center at Indiana U., p. 10.
- **RESEARCH WATCH:** *Humor in the Classroom*, James Rhem, Executive Editor, p. 10. What students find funny may not be politically correct, but it does seem to boost cognition.

## Making the Most of Exams: Procedures for Item Analysis

Raymond M. Zurawski, Ph.D.  
Associate Professor and Coordinator of  
Psychology  
St. Norbert College

One of the most important (if least appealing) tasks confronting faculty members is the evaluation of student performance. This task requires considerable skill, in part because it presents so many choices. Decisions must be made concerning the method, format, timing, and duration of the evaluative procedures. Once designed, the evaluative procedure must be administered and then scored, interpreted, and graded. Afterwards, feedback must be presented to students. Accomplishing these tasks demands a broad range of cognitive, technical, and interpersonal resources on the part of faculty. But an even more critical task remains, one that perhaps too few faculty undertake with sufficient skill and tenacity: investigating the **quality** of the evaluative procedure.

Even after an exam, how do we know whether that exam was a good one? It is obvious that any exam can only be as good as the items it comprises, but then *what constitutes a good exam item?* Our students seem to know, or at least *believe* they know. But are they

correct when they claim that an item was too difficult, too tricky, or too unfair?

Lewis Aiken (1997), the author of a leading textbook on the subject of psychological and educational assessment, contends that a "postmortem" evaluation is just as necessary in classroom testing as it is in medicine. Indeed, just such a postmortem

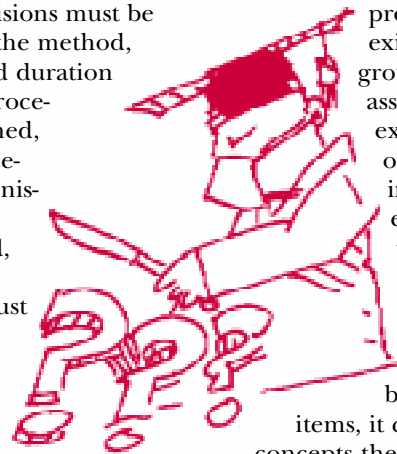
procedure for exams exists—*item analysis*, a group of procedures for assessing the quality of exam items. The purpose of an item analysis is to improve the quality of an exam by identifying items that are candidates for retention, revision, or removal. More specifically, not only can the item analysis identify both good and deficient

items, it can also clarify what concepts the examinees have and have not mastered.

So, what procedures are involved in an item analysis? The specific procedures involved vary, but generally, they fall into one of two broad categories: qualitative and quantitative.

### Qualitative Item Analysis

Qualitative item analysis procedures include careful proofreading of the exam prior to its administration for typographical errors, for grammatical cues that might inadvertently tip off examinees to the correct answer, and for the appropriateness of the reading level of the material.



Such procedures can also include small group discussions of the quality of the exam and its items with examinees who have already taken the test, or with departmental student assistants, or even experts in the field. Some faculty use a "think-aloud test administration" (cf. Cohen, Swerdlik, & Smith, 1992) in which examinees are asked to express verbally what they are thinking as they respond to each of the items on an exam. This procedure can assist the instructor in determining whether certain students (such as those who performed well or those who performed poorly on a previous exam) misinterpreted particular items, and it can help in determining *why* students may have misinterpreted a particular item.

### Quantitative Item Analysis

In addition to these and other qualitative procedures, a thorough item analysis also includes a number of quantitative procedures. Specifically, three numerical indicators are often derived during an item analysis: **item difficulty**, **item discrimination**, and **distractor power** statistics.

#### Item Difficulty Index ( $p$ )

The item difficulty statistic is an appropriate choice for achievement or aptitude tests when the items are scored dichotomously (i.e., correct vs. incorrect). Thus, it can be derived for true-false, multiple-choice, and matching items, and even for essay items, where the instructor can convert the range of possible point values into the categories "passing" and "failing."

The item difficulty index, symbolized  $p$ , can be computed simply by dividing the number of test takers who answered the item correctly by the total number of students who answered the item. As a proportion,  $p$  can range between 0.00, obtained when no examinees answered the item correctly, and 1.00, obtained when all examinees answered the item correctly. Notice that no test item need have only one  $p$  value. Not only may the  $p$  value vary with each class group that takes the test, an instructor may gain insight by computing the item difficulty level

for a number of different subgroups within a class, such as those who did well on the exam overall and those who performed more poorly.

Although the computation of the item difficulty index  $p$  is quite straightforward, the interpretation of this statistic is not. To illustrate, consider an item with a difficulty level of 0.20. We do know that 20% of the examinees answered the item correctly, but we cannot be certain *why* they did so. Does this item difficulty level mean that the item was challenging for all but the best prepared of the examinees? Does it mean that the instructor failed in his or her attempt to teach the concept assessed by the item? Does it mean that the students failed to learn the material? Does it mean that the item was poorly written? To answer these questions, we must rely on other item analysis procedures, both qualitative and quantitative ones.

#### Item Discrimination Index ( $D$ )

Item discrimination analysis deals with the fact that often different test takers will answer a test item in different ways. As such, it addresses questions of considerable interest to most faculty, such as, "does the test item differentiate those who *did well on the exam overall* from those who did not?" or "does the test item differentiate those who *know the material* from those who do not?" In a more technical sense then, item discrimination analysis addresses the validity of the items on a test, that is, the extent to which the items tap the attributes they were intended to assess. As with item difficulty, item discrimination analysis involves a family of techniques. Which one to use depends on the type of testing situation and the nature of the items. I'm going to look at only one of those, the *item discrimination index*, symbolized  $D$ . The index parallels the *difficulty index* in that it can be used whenever items can be scored dichotomously, as correct or incorrect, and hence it is most appropriate for true-false, multiple-choice, and matching items, and for those essay items which the instructor can score as "pass" or "fail."

We test because we want to find out if students know the material, but all we learn for certain is how

## THE NATIONAL TEACHING & LEARNING FORUM

### Executive Editor:

James Rhem, Ph.D.  
213 Potter St.  
Madison, WI 53715-2050

### Editorial Advisory Board

Jonathan Fife, Director Emeritus  
ERIC Clearinghouse on Higher Education

Judy Greene, Director  
Center for Teaching Effectiveness  
University of Delaware

Pat Hutchings, Senior Scholar  
The Carnegie Foundation for the Advancement  
of Teaching

Susan Kahn, Director  
AAHE Urban Universities Portfolio Project  
Indiana University-Purdue University

Wilbert McKeachie  
Professor of Psychology, Emeritus  
University of Michigan

Edward Neal, Director  
Center for Teaching and Learning  
University of North Carolina – Chapel Hill

Laura Rendón  
Professor of Education  
Arizona State University

Phyllis Steckler, President  
The Oryx Press

Marilla Svinicki  
Director, Center for Teaching Effectiveness  
University of Texas at Austin

### Editorial correspondence:

James Rhem  
213 Potter St.  
Madison, WI 53715-2050

### Subscription information:

The Oryx Press  
4041 North Central #700  
Phoenix, Arizona 85012  
Phone: 1-800-279-6799 or 602-265-2651

**The National Teaching & Learning Forum** (ISSN 1057-2880) is published six times during the academic year by The Oryx Press in conjunction with James Rhem & Associates, Inc. – October, December, February, March, May, September.

One-year individual subscription: \$39.

Periodicals postage paid at Phoenix, AZ.

Postmaster: Send change of address to:

**The National Teaching & Learning Forum**  
P.O. Box 33889  
Phoenix, Arizona 85067-3889

Copyright © 1998

James Rhem & Associates, Inc.

DUPLICATION BY PHOTOCOPYING OR  
OTHER MEANS IS STRICTLY FORBIDDEN.

**Bulk discount subscription rates available.**  
(Phone: 1-800-279-6799)

An E-mail edition of the Forum is \$32.00 for individuals. Significant discounts and site licenses available to entire institutions.

<http://www.ntlf.com>  
October

## Editor's Note:

The annoying qualities of all forms of mass communications remind us of what all communication asks of us. It asks that we give our attention and make a judgment. It seeks a relationship with us and we may not want another relationship. I do not want to think about the multi-leveled aspects of the Monica S. Lewinsky story, but in many ways, I have no choice. Or do I?

In the aftermath of the Starr Report being delivered to Congress, teachers all over the country began addressing the matter as a "teachable moment." A multitude of lessons stood ready to be discovered and learned. Could the *Forum* find any? Immediately I recalled an excellent case study by **David Brakke** that thoughtfully probes some of the dark and dangerous waters which surely lie in the human hearts behind the headlines. Need and longing, misunderstanding and gossip: these can affect faculty lives and cripple one's ability to teach. I've posted Dean Brakke's case on the *Forum's* Web site in the hope that you'll read it there, and begin there the thoughtful discussion the issues it raises deserve. Perhaps some good can come out of this mess.

It becomes clearer every day that we must take control of the Web and use it for our ends. Though paper and ink will never lose their sensual appeal and their utility, the Web will take over much of what paper has done. Those who lead in this wave of change have the chance to shape it. The University of Minnesota (where Gopher software was invented) sees ways of using the Web to amplify the conversation about teaching. The article on "webzines" by **Jan Smith** and **Paul Baeppler** shows how these early adopters of technology are using this publication and others to build a favorable teaching climate on their campus.

Perhaps good teaching and good communications are the same. How many times have we been bludgeoned with this insight? Its abuse lies in our knowledge that "communications" means so little until we break the idea open. A vital part of communications has to do with how we feel—how we feel about ourselves and how our students feel about themselves. Three articles in this issue get at this fact in different ways. **Laura Border's** DEVELOPER'S DIARY traces how feelings can trigger a learning crisis and powerful (and useful) insights into one's teaching. **David Tartakoff's** essay argues that students' self-esteem and confidence as learners transcend formally correct answers in teaching math. And my review of the research on using humor in the classroom suggests that, in some deep way, the brain is laughing when it is learning and that overt humor may help the process along.

The issue starts off, however, with an essay that seems quite different from all of these, **Raymond M. Zurawski's** discussion of what psychologists and statisticians call "item analysis." This article reminds readers that the exam isn't over when it's been graded. Exams themselves need questioning, and we can learn a lot about both the strengths and weaknesses of our teaching and of our students' learning from reviewing exam data in detail.

By the time you read this issue, over 25 campuses will have established site licenses to the *Forum*. That means more and more faculty will soon be reading the *Forum* on the Web. It's my hope some of them will want to discuss what they've read. To encourage such discussion, I've included links to the *Forum's* Web site discussion area. Those who don't read online can also post messages using this URL.

Together let us make the Web our own.

— James Rhem

they did on the exam we gave them. The item discrimination index tests the test in the hope of keeping the correlation between knowledge and exam performance as close as it can be in an admittedly imperfect system.

The item discrimination index is calculated in the following way:

1. Divide the group of test takers into two groups, high scoring and low scoring. Ordinarily, this is done by dividing the examinees into those scoring above and those scoring below the median. (Alternatively, one could create groups made up of the top and bottom quintiles or quartiles or even deciles.)

2. Compute the item difficulty levels separately for the upper ( $p_{upper}$ ) and lower ( $p_{lower}$ ) scoring groups.

3. Subtract the two difficulty levels such that  $D = p_{upper} - p_{lower}$ .

How is the item discrimination index *interpreted*? Unlike the item difficulty level  $p$ , the item discrimination index can take on negative values and can range between -1.00 and 1.00. Consider the following situation: suppose that overall, half of the examinees answered a particular item correctly, and that all of the examinees who scored above the median on the exam answered the item correctly and all of the examinees who scored below the median answered incorrectly. In such a situation  $p_{upper} = 1.00$  and  $p_{lower} = 0.00$ . As such, the value of the item discrimination index  $D$  is 1.00 and the item is said to be a perfect positive discriminator. Many would regard this outcome as ideal. It suggests that those who knew the material and were well-prepared passed the item while all others failed it.

Though it's not as unlikely as winning a million-dollar lottery, finding a perfect positive discriminator on an exam is relatively rare. Most psychometricians would say that items yielding positive discrimination index values of 0.30 and above are quite good discriminators and worthy of retention for future exams.

Finally, notice that the difficulty and discrimination are not independent. If all the students in both the upper and lower levels either pass or

fail an item, there's nothing in the data to indicate whether the item itself was good or not. Indeed, the value of the item discrimination index will be maximized when only half of the test takers overall answer an item correctly; that is, when  $p = 0.50$ . Once again, the ideal situation is one in which the half who passed the item were students who all did well on the exam overall.

Does this mean that it is never appropriate to retain items on an exam that are passed by all examinees, or by none of the examinees? Not at all. There are many reasons to include at least some such items. Very easy items can reflect the fact that some relatively straightforward concepts were taught well and mastered by all students. Similarly, an instructor may choose to include some very difficult items on an exam to challenge even the best-prepared students. The instructor should simply be aware that neither of these types of items functions well to make discriminations among those taking the test.

### Item Distractor Analysis

The final component of a good item analysis applies mainly to a particular type of test item, the multiple-choice item. On such items, the incorrect alternatives are called distractors. Item distractor analysis examines the percentage of examinees who select each incorrect alternative, to determine whether the distractors are functioning as intended.

To appreciate the logic of the analysis, consider Table 1 below. For each of four different items on a hypothetical multiple-choice exam, the table indicates the numbers of those in the upper and lower scoring groups on the overall exam

who selected each of the four possible response alternatives. On a well-designed multiple choice item (such as Item 1 shown below), those who know the material and are well-prepared for the exam should select the correct alternative even from among highly plausible distractors. Those who are not well-prepared should guess or select almost randomly from among the available

**We test because we want to find out if students know the material, but all we learn is how they did on the exam.**

distractors. Such an item would be a very good discriminator and would very likely be a *candidate for retention* for use in future exams, although it does have a relatively low level of difficulty overall.

Item distractor analysis can also provide useful diagnostic information in other situations. For example, in Item 2 in Table 1, the majority of those in the upper scoring group selected a wrong answer. This may indicate that the item has been inadvertently *miskeyed*, a proofreading error. Item 3 is a relatively difficult item overall; few examinees in either group answered it correctly. This item is a *candidate for revision*, particularly distractor *a*, which seems to be drawing undue attention as a plausible choice even from those who did well on the exam overall. Finally, Item 4 repre-

sents a *candidate for removal* from the exam. This item was passed by more of those who did poorly on the exam overall than those who were well-prepared and knew the material.

Could this data provoke other insights? Perhaps. It could be, for example, that the poor students did well because the teacher shifted learning styles that day, became more concrete, less abstract, and conveyed material more clearly to them than to the high-scoring, conceptual thoroughbreds.

### Conclusion

To those concerned about the prospect of extra work involved in item analysis, take heart: item difficulty and discrimination analysis programs are often included in the software used in processing exams answered on Scantron or other optically scannable forms. As such, these analyses can often be performed for you by personnel in your computer services office. You might consider enlisting the aid of your departmental student assistants to help with item distractor analysis, thus providing them with an excellent learning experience. In any case, an item analysis can certainly help determine whether or not the items on your exams were good ones and to determine which items to retain, revise, or replace. ■

### References:

- Aiken, L.R. (1997). *Psychological testing and assessment* (9th ed.). Boston, MA: Allyn and Bacon, Inc.
- Cohen, R.J., Swerdlik, M.E., & Smith, D.K. (1992). *Psychological testing and assessment: An introduction to tests and measurement* (2nd ed.). Mountain View, CA: Mayfield Publishing Company.

### Contact:

(Raymond Zurawski)  
zurarm@sncac.snc.edu

### Discussion:

[http://www.ntlf.com/ntlf\\_online/archive/](http://www.ntlf.com/ntlf_online/archive/)



**Table 1**  
Distractor Analysis Data for Items 1, 2, 3, and 4 on a Scoring Scale

Item	Upper Scoring Group				Lower Scoring Group			
	A	B	C	D	A	B	C	D
Item 1	10	0	0	0	10	10	10	10
Item 2	0	10	0	0	10	10	10	10
Item 3	0	0	0	0	0	0	0	0
Item 4	0	0	0	0	10	10	10	10

Note: C's are the correct alternative for each item. For each item, the total number of students in each group is 20.