Chapter 12.

Improving Test Items

Contents

- **Empirical And Judgmental Techniques**
- **Judgemental Improvement of Test Items**
- **Empirical Improvement of Test Items**

EMPIRICAL AND JUDGMENTAL TECHNIQUES

In considering the possible approaches to salvaging a flawed test item, it is helpful to isolate two fairly distinctive improvement strategies. The first of these is a judgmental approach in which the dominant reliance is on the human judgment that individuals render when they inspect, then weigh the merits of particular test items. A second approach to item improvement can be characterized as an empirical method.

Empirical Item Improvement: Improving an item's quality based on students' performance on the item.

Judgmental Item Iprovement: Improving an item's quality based on the options of student reviews.

JUDGEMENTAL IMPROVEMENT OF TEST ITEMS

Nonstudent reviewers

The sorts of review that might be employed by teachers themselves or external reviewers.

1. Is the item congruent with its assessment domain?
2. Are there violations of standard item-writing guidelines?
3. Is the content of the item accurate?
4. Is the item ethnically, socioeconomically, or otherwise biased?

1. Assessment domain congruence

To judge whether an item is congruent with the assessment domain from whence it supposedly sprang, judges obviously have to consider both the assessment domain and the items themselves.

Reviewers of items from tests aimed at criterion-referenced inferences will have to be highly attentive to the ingredients of a described assessment domain because item congruence is especially important.

2. Adherence to item-writing guidelines

A reviewer of items obviously needs to be familiar with a wide range of such rules.

If an item reviewer runs across a multiple-choice item in which the wrong-answer distractors are all short, while the correct answer is long, then that problem needs to be noted.

3. Content accuracy

If test items deal at all with academic content, such as achievement tests in history, language arts, of biology, then it is obviously important to have the content in those items be accurate.

4. Absence of bias

Item reviewers should be alert for blatant and subtle biases in items, whether racist, sexist, religious, gender, or socioeconomic.

Student judgment

Because students have experienced test items in a most meaningful, students judgment can provide useful insights regarding particular items and other features of the test such as its directions and the time allowed for completing the test.

1. If any of the items seemed confusing, which ones were they?
2. Did you think any items had more than one correct answer? If so, which ones?
3. Did you think any items had no correct answers? If so, which ones?
4. Were there words in any items that confused you? If so, which ones?
5. Were the directions for the test, or for particular subsections of the test, unclear? If so, which ones?

Student reviewers

Because students have experienced test items in a most meaningful manner, student judgment can provide useful insights regarding particular items and other features of the test such as its directions and the time allowed for completing the test.

1. If any of the items seemed confusing, which ones were they?
2. Did you think any items had more than one correct answer? If so, which ones?
3. Did you think any items had no correct answers? If so, which ones?
4. Were there words in any items that confused you? If so, which ones?
5. Were the directions for the test, or for particular subsections of the test, unclear? If so, which ones?

EMPIRICAL IMPROVEMENT OF TEST ITEMS

Difficulty Indices

Discrimination Indices

Distractor Analysis

Difficulty Indices

One useful index of an item's quality is its difficulty.

The most commonly employed item-difficulty index, often referred to these days simply as a p value, is calculated follows

$$\text{Difficulty } p = \frac{R}{T}$$

$$\text{Difficulty } p = \frac{37}{50} = .74$$

Difficulty Indices

P value

It should be clear that such p values can range from 0 to 1.00.

The p value of an item should be viewed in relationship to the student's chance probability of getting the correct response.

$$\text{Difficulty} = \frac{right}{total} X\,100$$

Difficulty Indices

A note of caution should be registered at this point, because measurement people sometimes err by referring to items with high p values, of .80 and above, as "easy" items, while items with low p values, of .20 and below, are described as "difficult" items. Those assertions may or may not be accurate.

Even though people typically refer to an item's p value as its difficulty index, the actual ease or difficulty of an item is almost always tied to the instructional program surrounding.

Difficulty index	Item evaluation
.00 ~ Less than .20	Very difficult item
.20 ~ .40	Difficult item
.40 ~ .60	Regular item
.60 ~ .80	Easy item
More than .80 ~ 1.00	Very easy item

Discrimination Indices

An item discrimination index typically indicates how frequently an item is answered correctly by those who perform well on the total test and how frequently an item is answered incorrectly by those who perform poorly on the total test.

An item discrimination index reflects the relationship between students' responses on the total test and their responses on a particular test item.

It is more important at norm-referenced evaluation than at criterion-referenced evaluation

One approach to computing an item-analysis statistic is to calculate a point biserial correlation coefficient between the continuous variable of total test score and the dichotomous variable of performance on a particular item.

Discrimination Indices

Type of item	Proportion of correct responses on total test
Positive Discriminator	High Scorers > Low Scorers
Negative Discriminator	High Scorers < Low Scorers
Nondiscriminator	High Scorers = Low Scorers

1. Order the test papers from high to low by total score.
2. Divide the papers into a high group and low group with an equal number of papers in each group.
3. Calculate a p value for each of the high and low groups.
4. Subtract p_l from p_h to obtain each item's discrimination index.

$$D = p_h - p_l$$

Discrimination Indices

Ebel's discrimination index values.

Discrimination index	Item evaluation
.40 and above	Very good items
.30 - .39	Reasonably good but possibly subject to improvement
.20 - .29	Marginal items, usually needing and being subject to improvement
.19 and below	Poor items, to be rejected or improved by revision

A review of the procedures for determining an item's discrimination index will suggest that an item's ability to discriminate is highly related to its overall difficulty index.

Distractor Analysis

In the case of multiple-choice items, teachers can gain further insights by carrying out a distractor analysis in which they see how the high and low groups are responding to the item's distractors.

Item No. 7	Alternatives				
(p =.50, D=-.33)	A	B*	C	D	Omit
Upper 16 students	2	5	0	8	1
Lower 15 students	4	10	0	0	1

When the Assessment Focus is on Criterion-reverenced interpretations

There are two general item-analysis schemes that have been employed thus far, depending on the kinds of criterion groups available. The first approach involves the administration of the test to the same group of students both prior to and following instruction. The second approach is to locate two different groups of students, one of whom has already been instructed and one of whom hasn't.

The same group of students :

A disadvantage of this approach is that one must wait for instruction to be completed before securing the item-analysis data.
Another problem is that the pretest may be reactive

When the Assessment Focus is on Criterion-reverenced interpretations

Two different groups of students :

It is possible to pick up some useful clues regarding item quality.
This approach has the advantage of avoiding the delay associated with pretesting and posttesting the same group of students and also of avoiding the possibility of a reactive pretest.

However, its drawback must rely on human judgment in the selection of the "instructed" and "uninstructed" groups.
The two groups should be identical in all other relevant respects.

When the Assessment Focus is on Criterion-reverenced interpretations

Pretest-posttest differences.

We can use an item discrimination index. This index is calculated as follows :

$$D_{ppd} = P_{post} - P_{pre}$$

Uninstructed versus instructed group differences.

If the item-improver uses two groups, an instructed and an uninstructed group, one of the more straight forward item discrimination indices is D_{uigd}.

$$D_{uigd} = P_i - P_u$$

Once Improved, the Number of Items

"How many test items should they actually use in creating the final version of their test?" is obviously an important question. Because if teachers use too few items in the test, they don't get an accurate fix on the students' status with respect to the assessment domain they're measuring. If teachers use too many items, there is lost economy on two counts : the unnecessary items they've produced and the unnecessary time taken from students as they wade through superfluous items.

Many factors unfortunately operate to confuse the situation so that no one can spin out a simple answer to the question of how many items.

Thank you.