Item Analysis with Small Samples

Baruch Nevo
University of Haifa, Israel

Traditional item analysis centers on the characteristics of individual items, typically on the item's level of difficulty and discrimination power. In constructing new tests, attempts are therefore made to obtain large samples of subjects in order to decrease the standard error of measurement of the item's characteristics. However, there are common test situations in which the exact parameters of individual items are not of much importance. Rather, the focus of interest is on the position of the items in relation to one another or in relation to some critical statistical value. Five such test situations are described. Quasi-simulations of item analyses were performed to determine the optimal sample sizes required in such test situations. These simulations consisted of analyzing responses of 5,200 university applicants, each of whom completed three different multiple-choice tests. Sample sizes of 16, 32, 64, 128, 256, 512, and 1,024 were chosen; and for each size, eight samples were randomly drawn from the population of applicants. For three of five different indices of accuracy that were employed, the results showed that the sample size needed for the pretest stage in test construction is considerably smaller than the traditionally recommended size.

Traditional applications of item analysis require samples of several hundred subjects. When developing a new test, or pretesting items for a new form of an existing test, the test constructor would try to obtain a large sample; in certain institutions the size of such a sample may be greater than 1,000. Why large samples? The answer is obvious: The larger the sample, the smaller the standard error of the items' characteristics. The index of difficulty of an item in the population measured by the percentage of correct responses (P), the item-total score correlation in the population (ϱ), and other items' parameters can be estimated more accurately when a larger sample is employed.

Following this kind of logic, several test specialists calculated the minimum required number of subjects for a pretest. Henryson (1971), for instance, recommended a minimum of 300 testees. Nunnally (1967) stated that "all items should be administered to at least 300 persons, preferably to 1,000 or more" (p. 242). Conrad's (1948) minimum is 500. At Educational Testing Service, test developers employ 1,000 to 3,000 subjects when pretesting new items for the Scholastic Aptitude Test (Swineford, 1974).

In a situation in which a shorter confidence interval of a parameter of an *individual* item is needed, a large sample is probably a necessity that cannot be avoided. This is, for example, the situation when a bank of items is under construction. However, there are certain situations in which the parameters of individual items are *not* of great importance, and a pretest of items could be carried out, presumably with smaller

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 3 Summer 1980 pp. 323–329

samples. These situations and the sizes of samples they require are the main concern of this study. Five such situations will be described here.

1. As an example of a situation in which the exact value of an individual item's characteristic is not of great importance, consider the case of a test constructor who is pretesting a group of *m* items and whose purpose is to *order* them according to their *P*'s. Sampling errors here should be measured by comparing the population's order of the items' *P*'s to the sample's order of *p*'s. For instance, if $m = 4$ and the population levels of item difficulty are $P = .20$, .40, .60, and .80 for Items A, B, C, and D, respectively, then a sample's data of $p = .22$, .33, .88, and .90 would lead the test constructor to the correct order. The sample's results could be considered, therefore, very satisfactory, even though an item by item comparison reveals large discrepancies.
2. In the second situation, the test constructor wishes to order the items according to the levels of their ϱ's. This procedure is useful when, for instance, out of *m* pretested items, only *n* ($n < m$) with the highest ϱ's are to be chosen for the final form.
3. As an example of the third situation, consider a test constructor who wishes to discard all the items with $\varrho > .30$. When pretesting a pool of items, he/she is not really interested in the exact value of the true correlation of each item, only in whether this value is above or below the critical value (which is .30 in this example). Accordingly, the accuracy of the sample statistics is measured here, not in terms of the distance between statistic and parameter of an item, but in terms of the proportion of consistent classifications of the items (rate of concordance) in the pretest pool. An item will be misclassified if, for instance, $\varrho = .35$ and its correlation in the pretest sample (*r*) is .28. On the other hand, an item with true (population value) correlation of $\varrho = .55$ that appears in the pretest sample as having a correlation of $r = .75$ will not be misclassified.
4. In the fourth situation the relevant item characteristic is, again, the level of difficulty, but the context is different from that of the first situation. The test constructor is interested here in producing a test consisting of *n* items whose *P*'s fit a specified distribution (e.g., rectangular, normal). He/she would pretest *m* items from which *n* will be selected to fit the required frequency distribution. This selection must be based on the items' *p*'s, since their *P*'s are found only after the test is administered to the population. The selection could be considered successful if the *P*'s of the *n* items form the required distribution. A difference between *p* and *P* of any individual item has no importance because it may be nullified by a reversed difference in another item.
5. The fifth situation is a combination of the third and fourth situations. The test constructor is pretesting *m* items, from which he/she will select *n* so that *each one* of the *items* has a higher-than-critical value of ϱ and as *a group* their *P*'s are distributed according to a prespecified model.

In these five situations the focus of interest is not on the individual item but on the position of the items in relation to each other or in relation to some critical value. The situations were chosen because they are quite common among "small caliber" test constructors around the globe. Typically, a test is needed for a limited use, and the resources for experimenting with items are confined. Frequently, there will be some difficulty in getting several thousands or even several hundreds of subjects for the pretest phase. It seems quite important to know what kinds of accuracy losses are involved when a small sample is employed. By having this kind of information, the test constructor will be in a better position to decide what kind of "compromise" regarding sample size is reasonable. The present study is intended to gather *empirical* information regarding the following questions: (1) How much accuracy is lost in each situation? and (2) at what *pace* is it happening when the size of the sample is gradually reduced?

Is it possible to develop a theoretical model for the sampling distributions of accuracy indices defined for the above-mentioned situations? This is an important question because if the answer is positive, much time and money can be saved by avoiding an empirical data analysis. However, even though in principle such a model might be possible, it will certainly be very difficult to establish. The parameters of the sampling distribution of any such index will depend on at least three factors: the number of items in the test, the sample size, and the parameters of the individual items. A theoretical model should combine within its framework all three factors; otherwise, it will not work. Some of Levy's recent works might be used as potential starting points (Levy, 1975, 1976, 1977), but at present none of them is directly applicable to the theoretical questions presented here. On the other hand, there are many examples of the usefulness of an empirical approach—an approach followed in this study.

Method

Tests

Three tests were utilized: Vocabulary (30 items), Arithemetical Reasoning (35 items), and an English mastery test (45 items). All three tests were part of the National Entrance Examination battery administered to applicants to universities in Israel in 1977. The tests consisted of multiple-choice items only. The difficulty indices of the items (P's) and their item-total score point-biserial correlations (ϱ's) were calculated; they created the basis for further analysis.

Population and Samples

The above-mentioned battery was taken by 5,200 subjects. Random samples were drawn with replacement from this population. The sizes of the samples were 16, 32, 64, 128, 256, 512, and 1,024. For each of these seven sample sizes, eight different samples were drawn, for a total of 56 samples. For each item in each sample, the difficulty index (p) and the item-total score point-biserial correlation (r) were calculated.

Indices of Accuracy

Five accuracy indices were defined, each of which is associated with one of the situations described above. These indices were measures of the goodness of fit of a sample statistic to the population parameters.

I_1: This index of accuracy was a Spearman rank-order correlation between the order of the items' P's (the items' level of difficulty in the population) and the order of the sample p's of the same items.

I_2: This index was the Spearman rank-order correlation between the order of the items' ϱ's (the population item-total score point-biserial correlation) and the order of the sample r's of the same items.

I_3: For each test the median value of the items' ϱ's was found. The medians for the three tests were .43 for Vocabulary, .41 for Arithmetical Reasoning, and .39 for English. The median served (arbitrarily) as a critical value for the classification of the r's. An item was counted as a "consistent classification" if both its r and ϱ were *above* or below the critical value. I_3 was the proportion of the items consistently classified.

I_4: For each test, out of its m items, n were selected in such a way that their p's would create a rectangular distribution. The selection ratio, $n{:}m$, was about 1:2. The shape of the distribution and the items' selection ratio were arbitrarily chosen. This procedure was repeated with each sample and with the population's P's as well. I_4 was the proportion of items selected both by the population and sample data.

I_5: The fifth index was similar to I_4 except that when selecting the n items, not only were the items' P's (or p's) taken into consideration but

Table 1

Medians and Lower Quartiles of Five Indices of Accuracy of Item Analysis in Seven Sample Sizes for Three Tests

Index and Sample Size	Vocabulary (30 Items)		Arithmetical Reasoning (35 Items)		English (45 Items)	
	Q_1	Md	Q_1	Md	Q_1	Md
Index I_1						
16	.73	.75	.80	.82	.81	.82
32	.92	.93	.92	.94	.87	.90
64	.97	.98	.96	.96	.93	.94
128	.99	.99	.98	.98	.95	.95
256	.99	.99	.99	.99	.98	.98
512	1.00	1.00	.99	.99	.99	.99
1024	1.00	1.00	1.00	1.00	1.00	1.00
Index I_2						
16	.39	.55	.92	.47	.47	.56
32	.64	.69	.55	.65	.60	.66
64	.80	.85	.71	.74	.74	.78
128	.91	.95	.84	.88	.89	.91
256	.96	.96	.92	.93	.90	.93
512	.98	.98	.95	.96	.96	.97
1024	.99	.99	.96	.96	.99	.99
Index I_3						
16	.58	.68	.61	.66	.69	.75
32	.67	.72	.66	.70	.67	.74
64	.77	.83	.81	.86	.75	.76
128	.88	.92	.86	.91	.87	.89
256	.93	.95	.88	.91	.90	.92
512	.93	.93	.91	.93	.92	.93
1024	.97	.97	.94	1.00	.97	.98

Note: I_1 and I_2 are rank order correlations while I_3, I_4 and I_5 are concordance rates.

(continued on next page)

Table 1 (continued)

Medians and Lower Quartiles of Five Indices of Accuracy of Item Analysis in Seven Sample Sizes for Three Tests

Index and Sample Size	Vocabulary (30 Items)		Arithmetical Reasoning (35 Items)		English (45 Items)	
	Q_1	Md	Q_1	Md	Q_1	Md
Index I_4						
16	.73	.77	.72	.78	.73	.78
32	.73	.77	.78	.83	.76	.78
64	.73	.80	.72	.77	.76	.80
128	.77	.83	.81	.83	.76	.78
256	.80	.87	.78	.83	.80	.83
512	.90	1.00	.89	.89	.82	.85
1024	.97	1.00	.83	.83	.85	.89
Index I_5						
16	.67	.70	.64	.69	.56	.63
32	.70	.73	.67	.75	.63	.70
64	.73	.77	.69	.72	.67	.74
128	.73	.83	.78	.83	.70	.76
256	.80	.80	.78	.78	.78	.82
512	.90	1.00	.87	.89	.87	.87
1024	.97	1.00	.72	.72	.87	.96

also their ϱ's (or r's): Only items with ϱ (or r) higher than .35 were selected.

Indices I_1 and I_2 were rank-order correlations by their nature, whereas I_3, I_4, and I_5 were concordance rates.

Results

I_1 to I_5 were calculated for each sample. Since there were eight samples for every category of sample size, the median (Md) and the lower quartile (Q_1) values of each index were found for each category. Table 1 presents these values.

As might be expected, all five indices were found to be positively and monotonically related to the sample size: The larger the sample size, the higher the value of its indices. Exceptions to this rule were rare and should be related to sampling errors (i.e., I_3 Vocabulary equaled .95 for sample size of 256 and only .93 for sample size of 512). Beyond this generalization, which is a trivial one, it seems that the indices' values depended not only on the sample size but also on the specific test and specific type of index. The Vocabulary indices were, in general, higher than the indices of the other two tests. When the distributions of the P's and ϱ's of the three tests were compared, it was found that the Vocabulary distributions were slightly more heterogeneous than the other two (data are not presented here). This difference is hypothesized to be the major reason for the differences between the tests' indices.

When the five indices were compared to each other, I_1 approached its asymptotic value more rapidly than any other indice. This was true for Md and Q_1 as well. It seems that a sample of about 100 subjects is adequate for the purpose of ordering the items according to their "true" level of difficulty if a value of I_1 .90 is (arbitrarily) chosen to be satisfactory. Q_1 of I_2 and I_3 also reached a reasonable level with samples of 100 to 200 subjects. This means that a test constructor can order the items according to their expected ϱ's (I_2) or can select items whose expected ϱ is higher than a critical value (I_3) with relatively small samples. A report by Brogden (1956) essentially supports the findings regarding I_1 and I_2 (I_3 was not employed). If the test constructor is interested, however, in producing a specific and exact shape of distribution of items' P's (I_4) or ϱ's (I_5), and if he/she is working with tests similar to those employed in this study, he/she needs at least 500 to 1,000 subjects.

Md figures are higher than Q_1 in almost every cell of Table 1. If a test constructor refers to the Md figures (rather than the Q_1's) when deciding about the required sample size, he/she may arrive at smaller numbers; but this policy seems too hazardous because it leaves 50% chance of having less than optimal items' selection or items' ordering.

Discussion and Limitations

One major problem in this study was introduced by the fact that the tests that were used were *not* in an experimental stage: They were in their final operational form, and the procedures described above were *simulations* of item analyses. As a result, the ϱ's are high (because the items of the tests were preselected in that direction and items with low ϱ's were already discarded) and were similar to each other. Unfortunately, it is not easy to find an experimental test that has been administered to a "population," or to a large number of testees. It is reasonable, however, to assume that with real experimental tests the accuracy indices will be higher than those presented in Table 1; and the samples required, smaller.

Without a general model (see the introduction section) there is a serious problem of generalization from present findings. The results of this study cannot be applied directly to tests or to questionnaires that greatly differ in their qualities from those employed here. Much more data need to be gathered before any conclusive recommendations regarding item analysis can be made.

References

Brogden, H. E. Implications of item-index intercorrelations for item analysis. (Technical Research Note No. 62). U.S. Army, Personnel Research Branch, 1956.

Conrad, S. H. Characteristics and uses of item-analysis data. *Psychological Monographs*, 1948, *62*, 1-48.

Henryson, S. Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.), Washington, DC: Council on Education, 1971.

Levy, K. L. Appropriate sample sizes for selecting the population with the largest value of *r* from among *K* binomial populations. *British Journal of Mathematical and Statistical Psychology*, 1975, *28*, 134-137.

Levy, K. L. Testing for a priori trends in *K* independent correlations. *Educational and Psychological Measurement*, 1976, *36*, 671-674.

Levy, K. L. Appropriate sample sizes for selecting a population with the largest correlation coefficient from among *K* bivariate normal populations. *Educational and Psychological Measurement*, 177, *37*, 61-66.

Nunnally, J. C. *Psychometric theory*. New York: McGraw-Hill, 1967.

Swineford, F. *The test consultant manual*. Princeton, NJ: Educational Testing Service, 1974.

Acknowledgment

This study was supported by a grant from the Ford Foundation received through the Israel Foundations Trustees.

Author's Address

Send requests for reprints or further information to Baruch Nevo, Department of Psychology, University of Haifa, Haifa, Israel.